UNIT I

Measures of Central Tendency

Introduction

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

Mean (Arithmetic)

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data (see our Types of Variable guide for data types). The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values x_1, x_2, ..., x_n, the sample mean, usually denoted by $\bar{x}$ (pronounced x bar), is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

This formula is usually written in a slightly different manner using the Greek capitol letter, $\sum$, pronounced "sigma", which means "sum of...":

$$\bar{x} = \frac{\sum x}{n}$$

You may have noticed that the above formula refers to the sample mean. So, why have we called it a sample mean? This is because, in statistics, samples and populations have very different meanings and these differences are very important, even if, in the case of the mean, they are calculated in the same way. To acknowledge that we are calculating the population mean and not the sample mean, we use the Greek lower case letter "mu", denoted as µ:

$$\mu = \frac{\sum x}{n}$$

The mean is essentially a model of your data set. It is the value that is most common. You will notice, however, that the mean is not often one of the actual values that you have observed in your data set. However, one of its important properties is that it minimises error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set.

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

When not to use the mean

The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below:

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

The mean salary for these ten staff is $30.7k. However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the $12k to 18k range. The mean is being skewed by the two large salaries. Therefore, in this situation, we would like to have a better measure of central tendency. As we will find out later, taking the median would be a better measure of central tendency in this situation.

Another time when we usually prefer the median over the mean (or mode) is when our data is skewed (i.e., the frequency distribution for our data is skewed). If we consider the normal distribution - as this is the most frequently assessed in statistics - when the data is perfectly normal, the mean, median and mode are identical. Moreover, they all represent the most typical value in the data set. However, as the data becomes skewed the mean loses its ability to provide the best central location for the data because the skewed data is dragging it away from the typical value. However, the median best retains this position and is not as strongly influenced by the skewed values. This is explained in more detail in the skewed distribution section later in this guide.

Median

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

65	55	89	56	35	14	56	55	87	45	92

We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	**56**	56	65	87	89	92

Our median mark is the middle mark - in this case, 56 (highlighted in bold). It is the middle mark because there are 5 scores before it and 5 scores after it. This works fine when you have an odd number of scores, but what happens when you have an even number of scores? What if you had only 10 scores? Well, you simply have to take the middle two scores and average the result. So, if we look at the example below:

65	55	89	56	35	14	56	55	87	45

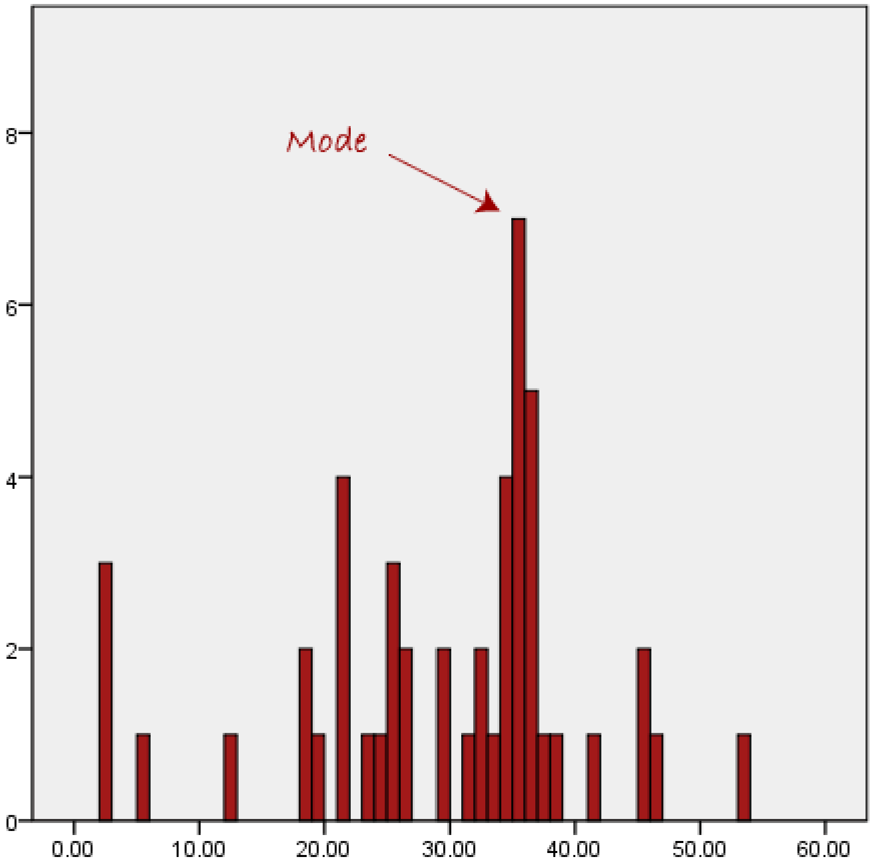
We again rearrange that data into order of magnitude (smallest first):

14	35	45	55	**55**	**56**	56	65	87	89

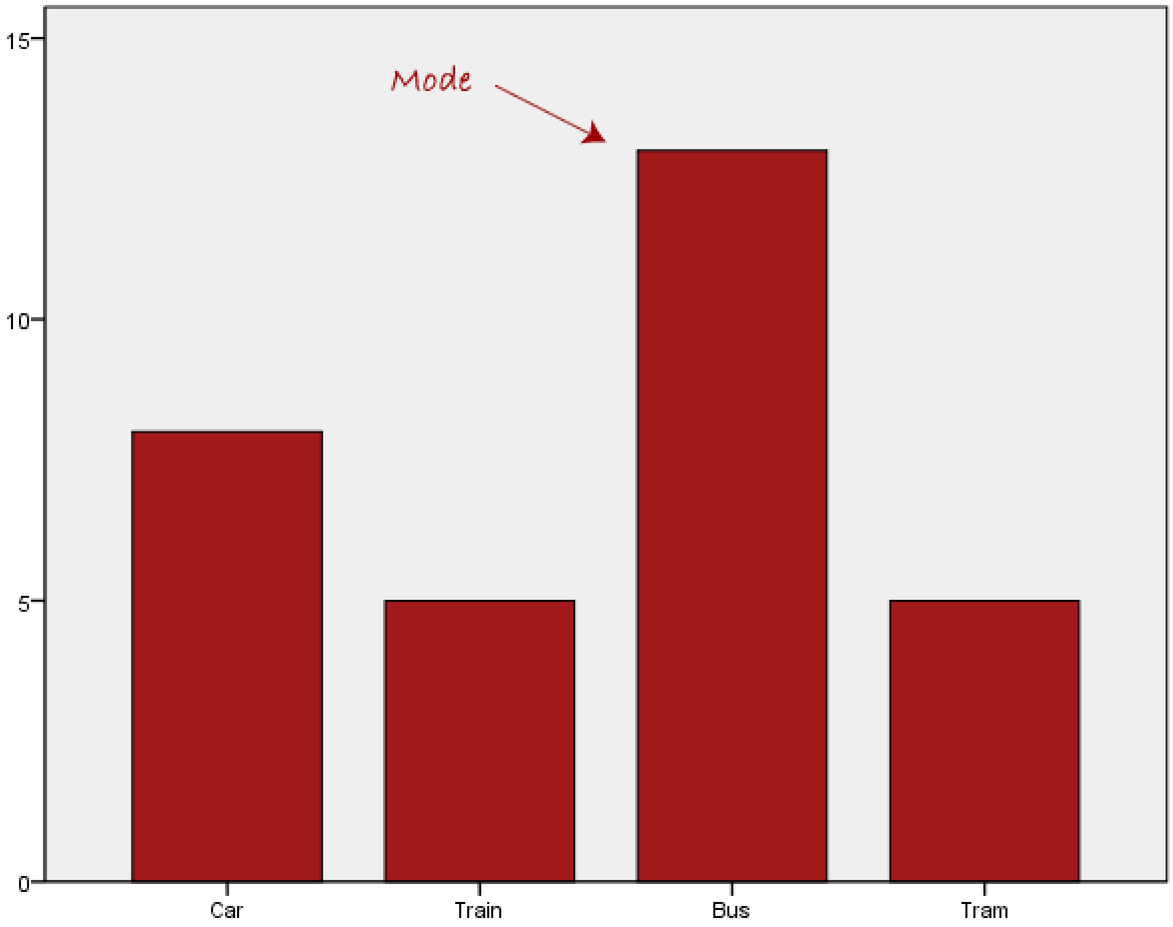
Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

Mode

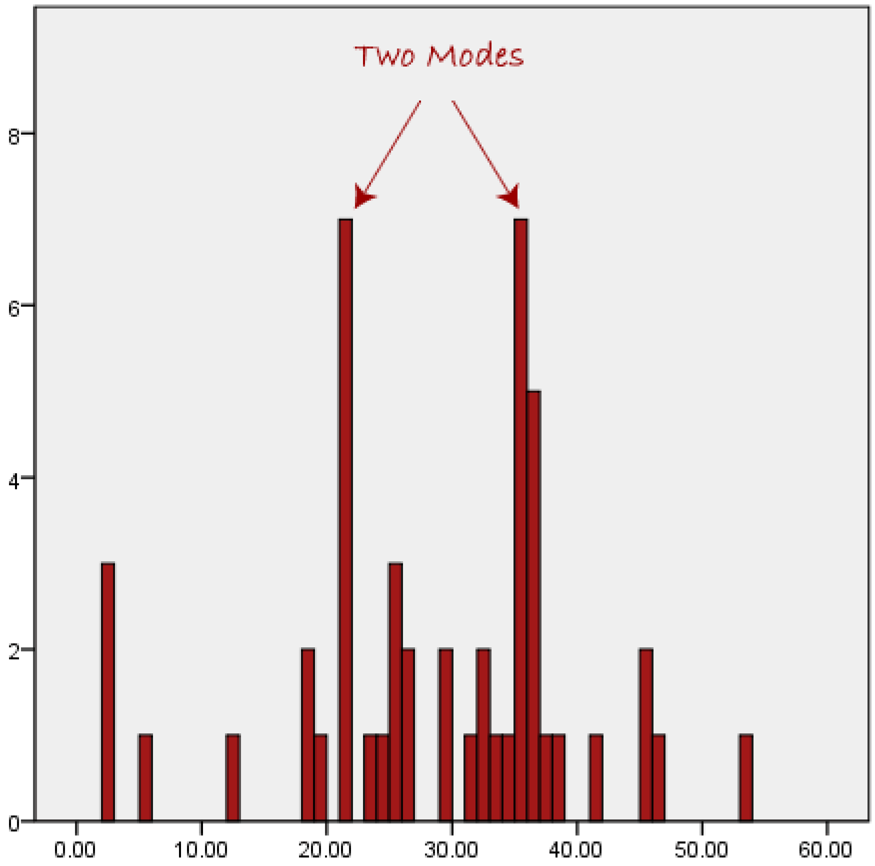
The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option. An example of a mode is presented below:



Normally, the mode is used for categorical data where we wish to know which is the most common category, as illustrated below:

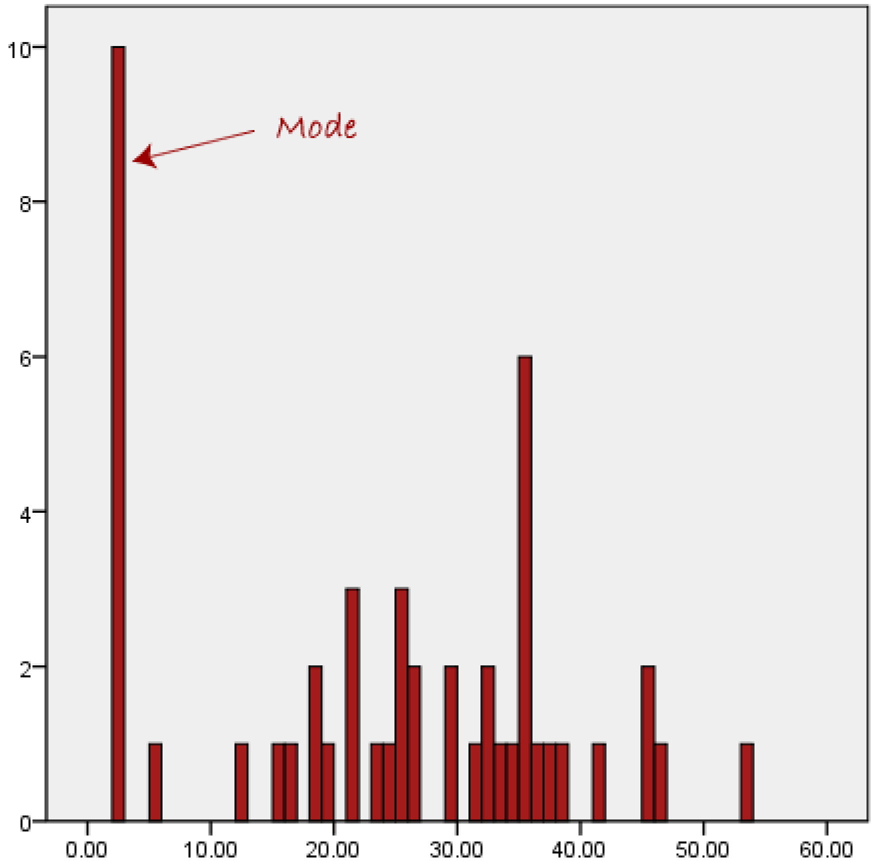


We can see above that the most common form of transport, in this particular data set, is the bus. However, one of the problems with the mode is that it is not unique, so it leaves us with problems when we have two or more values that share the highest frequency, such as below:



We are now stuck as to which mode best describes the central tendency of the data. This is particularly problematic when we have continuous data because we are more likely not to have any one value that is more frequent than the other. For example, consider measuring 30 peoples' weight (to the nearest 0.1 kg). How likely is it that we will find two or more people with **exactly** the same weight (e.g., 67.4 kg)? The answer, is probably very unlikely - many people might be close, but with such a small sample (30 people) and a large range of possible weights, you are unlikely to find two people with exactly the same weight; that is, to the nearest 0.1 kg. This is why the mode is very rarely used with continuous data.

Another problem with the mode is that it will not provide us with a very good measure of central tendency when the most common mark is far away from the rest of the data in the data set, as depicted in the diagram below:



In the above diagram the mode has a value of 2. We can clearly see, however, that the mode is not representative of the data, which is mostly concentrated around the 20 to 30 value range. To use the mode to describe the central tendency of this data set would be misleading.

Harmonic Mean Definition:

Harmonic mean is used to calculate the average of a set of numbers. Here the number of elements will be averaged and divided by the sum of the reciprocals of the elements. The Harmonic mean is always the lowest mean.

Harmonic Mean Formula :

Harmonic Mean = $N/(1/a_1+1/a_2+1/a_3+1/a_4+.......+1/a_N)$ **Where,**

X = Individual score N = Sample size (Number of scores)

Harmonic Mean Example:

To find the Harmonic Mean of 1,2,3,4,5.

Step 1:

Calculate the total number of values. N = 5

Step 2:

Now find Harmonic Mean using the above formula. $N/(1/a_1+1/a_2+1/a_3+1/a_4+.......+1/a_N)$ = 5/(1/1+1/2+1/3+1/4+1/5) = 5/(1+0.5+0.33+0.25+0.2) = 5/2.28 So, Harmonic Mean = 2.19 This example will guide you to calculate the harmonic mean manually.

Geometric Mean Definition:

Geometric mean is a kind of average of a set of numbers that is different from the arithmetic average. The geometric mean is well defined only for sets of positive real numbers. This is calculated by multiplying all the numbers (call the number of numbers n), and taking the nth root of the total. A common example of where the geometric mean is the correct choice is when averaging growth rates.

Formula:

Geometric Mean :

Geometric Mean = $((X_1)(X_2)(X_3)........(X_N))^{1/N}$

where

X = Individual score N = Sample size (Number of scores)

Geometric Mean Example:

To find the Geometric Mean of 1,2,3,4,5.

Step 1:

N = 5, the total number of values. Find 1/N. 1/N = 0.2

Step 2:

Now find Geometric Mean using the formula. $((1)(2)(3)(4)(5))^{0.2} = (120)^{0.2}$ So, Geometric Mean = 2.60517 This example will guide you to calculate the geometric mean manually.

Combined Mean

Example

	Series A	Series A
Mean	50	40
Standard deviation	5	6
No. of items	100	150

Find the combined mean of the two series.

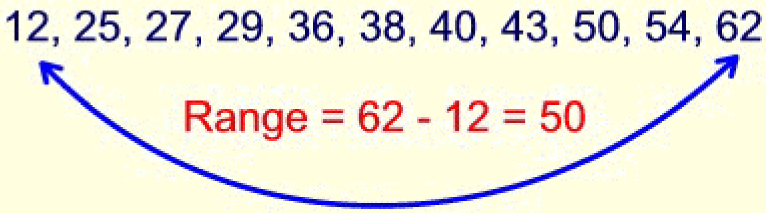
Solution

$\bar{X}_{12} = N_1X_1 + N_2\, X_2 / N_1 + N_2$

= (100×50)+ (150×40))/(100+150)=11,000/250=44

UNIT II

Range: The simplest of our methods for measuring dispersion is *range*. Range is the difference between the largest value and the smallest value in the data set. While being simple to compute, the range is often unreliable as a measure of dispersion since it is based on only two values in the set.



12, 25, 27, 29, 36, 38, 40, 43, 50, 54, 62

Range = 62 - 12 = 50

A range of 50 tells us very little about how the values are dispersed.
Are the values all clustered to one end with the low value (12) or the high value (62) being an outlier?
Or are the values more evenly dispersed among the range?

Standard Deviation: Standard deviation is the square root of the variance. The formulas are:

$$\text{Population standard deviation} = \sigma x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\text{Sample standard deviation} = Sx = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Variance: To find the variance:

- subtract the mean, $\overline{x}$, from each of the values in the data set, x_i.
- square the result
- add all of these squares
- and divide by the number of values in the data set.

$$\text{Population variance} = (\sigma x)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\text{Sample variance} = (Sx)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Combined Standard deviation

Example

	Series A	Series A
Mean	50	40
Standard deviation	5	6
No. of items	100	150

Find the combined standard deviation of the two series.

$\sigma_{12} = \sqrt{}\ N_1(\sigma_1^2 + d_1^2) + N_2\,(\sigma_2^2 + d_2^2)\ /\ N_{1+\ N2}$

$d_1 = (\bar{X}_1 - \bar{X}_{12}) = 50 - 44 = 6$

$d_2 = (\bar{X}_2 - \bar{X}_{12}) = 40 - 44 = 4$

$\sigma_{12} = \sqrt{}(100\ [\ (5_2 + 6^2)\ ] + 150\ [6^2 + (-4^2\ )])\ /\ (100+150)))$

$= \sqrt{}((2500+3600+5400+2400)/250)\ \ = \sqrt{}(13{,}900/250)$

$= \sqrt{}55.6 = 7.46$

Coefficient of Variation

A coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. It is calculated as follows:

$$\text{Coefficient of Variation} = \frac{\textbf{Standard Deviation}}{\textbf{Expected Return}}$$

The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other.

UNIT III

In correlation, when values of one variable increase with the increase in another variable, it is supposed to be a **positive correlation**. On the other hand, if the values of one variable decrease with the decrease in another variable, then it would be a **negative correlation**. There might be the case when there is no change in a variable with any change in another variable. In this case, it is defined as **no correlation** between the two.

Correlation can be of three types as follows:

1. Simple correlation
2. Multiple correlation
3. Partial correlation

Correlation Definition

The relationship between more than one variable is considered as correlation. Correlation is considered as a number which can be used to describe the relationship between two variables. Simple correlation is defined as a variation related amongst any two variables. The multiple correlation and partial correlation are categorized as related variation among three or more variables. Two variables are correlated only when they vary in such a way that the higher and lower values of one variable corresponds to the higher and lower values of the other variable. We might also get to know if they are correlated when the higher value of one variable corresponds with the lower value of the other.

Correlation Symbol

Symbol of correlation = r

Correlation Formula

The formula for correlation is as follows,

Correlation (r)
$= N\sum XY-(\sum X)(\sum Y)[N\sum X_2-(\sum X)_2][N\sum Y_2-(\sum Y)_2]\surd N\sum XY-(\sum X)(\sum Y)[N\sum X2-(\sum X)2][N\sum Y2-(\sum Y)2]$

Where,
x and y are the variables.
b = the slope of the regression line is also called as the regression coefficient
a = intercept point of the regression line which is in the y-axis.
N = Number of values or elements
X = First Score
Y = Second Score

$\sum XY$∑XY = Sum of the product of the first and Second Scores
$\sum X$∑X = Sum of First Scores
$\sum Y$∑Y = Sum of Second Scores
$\sum X2$∑X2 = Sum of square first scores.
$\sum Y2$∑Y2 = Sum of square second scores.

Coefficient of Correlation

Coefficient of correlation, r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. It also called as Pearson product moment correlation coefficient. The algebraic method of measuring the correlation is called the coefficient of correlation. There are mainly three coefficients of correlation

1. Karl Pearson’s Coefficient of correlation
2. Pearson’s rank correlation coefficient
3. Concurrent correlation

Interpretation of Karl Pearson’s Coefficient of correlation

Karl Pearson’s Coefficient of correlation denoted by r is the degree of correlation between two variables. r takes values between –1 and 1

When r is –1, we say there is perfect negative correlation.
When r is a value between –1 and 0, we say that there is a negative correlation
When r is 0, we say there is no correlation
When r is a value between 0 and 1, we say there is a positive correlation
When r is 1, we say there is a perfect positive correlation.

Properties of the Coefficient of correlation

1. Coefficient of correlation has a well defined formula
2. Coefficient of correlation is a number and is independent of the unit of measurement
3. Coefficient of correlation lies between –1 and 1
4. Coefficient of correlation between x and y will be same as that between y and x.

Types of Correlation

There are different types of Correlation. They are listed as follows:

Positive Correlation

A positive correlation is a correlation in the same direction.

Negative Correlation

A negative correlation is a correlation in the opposite direction.

Partial Correlation

The correlation is partial if we study the relationship between two variables keeping all other variables constant.

Example:

The Relationship between yield and rainfall at a constant temperature is partial correlation.

Linear Correlation

When the change in one variable results in the constant change in the other variable, we say the correlation is linear. When there is a linear correlation, the points plotted will be in a straight line

Example:

Consider the variables with the following values.

X:	10	20	30	40	50
Y:	20	40	60	80	100

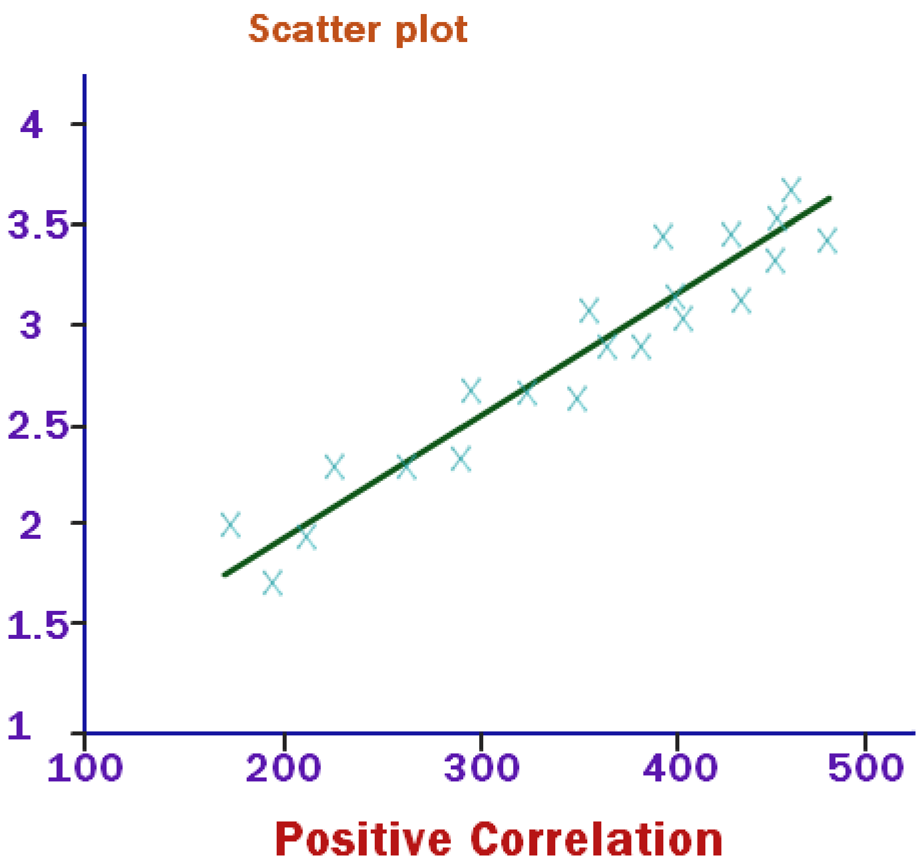
Here, there is a linear relationship between the variables. There is a ratio 1:2 at all points. Also, if we plot them they will be in a straight line.

Zero Order Correlation

One of the most common and basic techniques for analyzing the relationships between variables is zero-order correlation. The value of a correlation coefficient can vary from -1 to +1. A -1 indicates a perfect negative correlation, while a +1 indicates a perfect positive correlation. A correlation of zero means there is no relationship between the two variables.

Scatter Plot Correlation

A scatter plot is a type of mathematical diagram using cartesian coordinates to display values for two variables for a set of data. Scatter plots will often show at a glance whether a relationship exists between two sets of data. The data displayed on the graph resembles a line rising from left to right. Since the slope of the line is positive, there is a positive correlation between the two sets of data.



Spearman's Correlation

Spearman's rank correlation coefficient allows us to identify easily the strength of correlation within a data set of two variables, and whether the correlation is positive or negative. The Spearman coefficient is denoted with the Greek letter rho (ρ).

Non Linear Correlation

When the amount of change in one variable is not in a constant ratio to the change in the other variable, we say that the correlation is non linear.

Example:

Consider the variables with the following values

X:	10	20	30	40	50
Y:	10	30	70	90	120

Here there is a non linear relationship between the variables. The ratio between them is not fixed for all points. Also if we plot them on the graph, the points will not be in a straight line. It will be a curve.

Non linear correlation is also known as curvilinear correlation.

Simple Correlation

If there are only two variable under study, the correlation is said to be simple.

Example:

The correlation between price and demand is simple.

Multiple Correlations

When one variable is related to a number of other variables, the correlation is not simple. It is multiple if there is one variable on one side and a set of variables on the other side.

Example:

Relationship between yield with both rainfall and fertilizer together is multiple correlations

Weak Correlation

The range of the correlation coefficient between -1 to +1. If the linear correlation coefficient takes values close to 0, the correlation is weak.

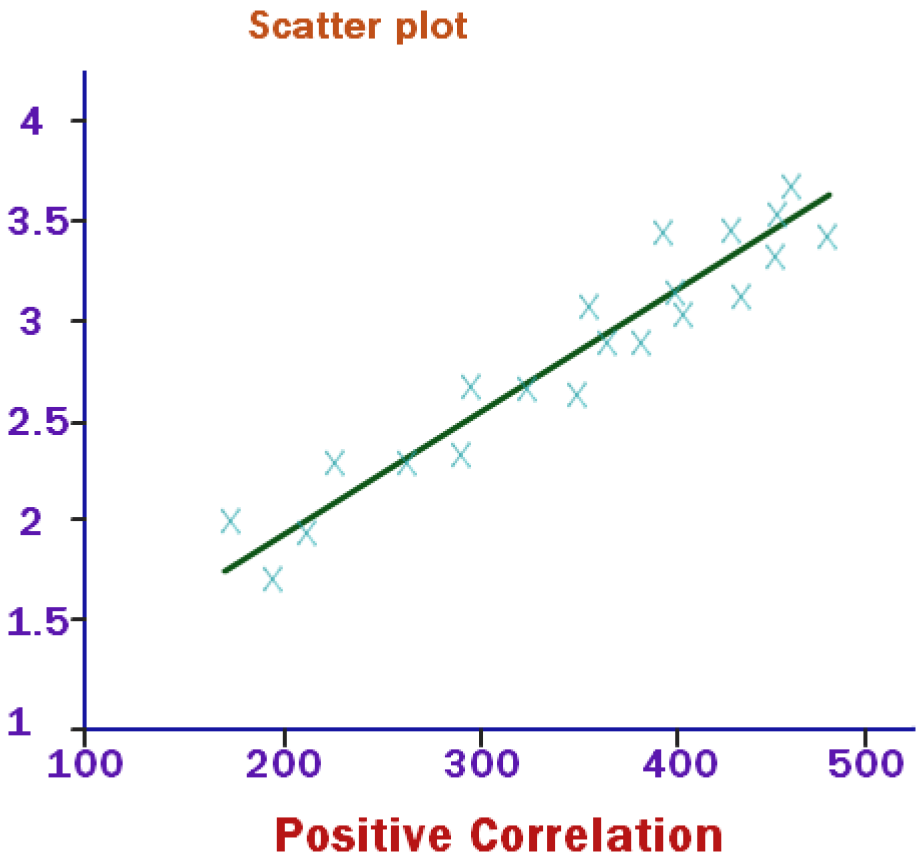
Positive Correlation

A relationship between two variables in which both variables move in same directions. A positive correlation exists when as one variable decreases, the other variable also decreases and vice versa. When the values of two variables x and y move in the same direction, the correlation is said to be **positive**. That is in positive correlation, when there is an increase in x, there will be and an increase in y also. Similarly when there is a decrease in x, there will be a decrease in y also.

Positive Correlation Example

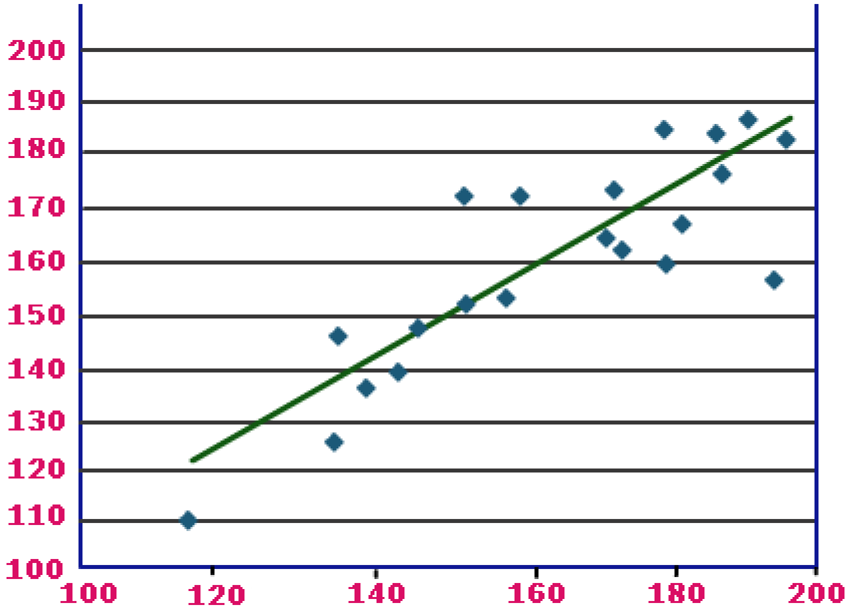
Price and supply are two variables, which are positively correlated. When Price increases, supply also increases; when price decreases, supply decreases.

Positive Correlation Graph



Strong Positive Correlation

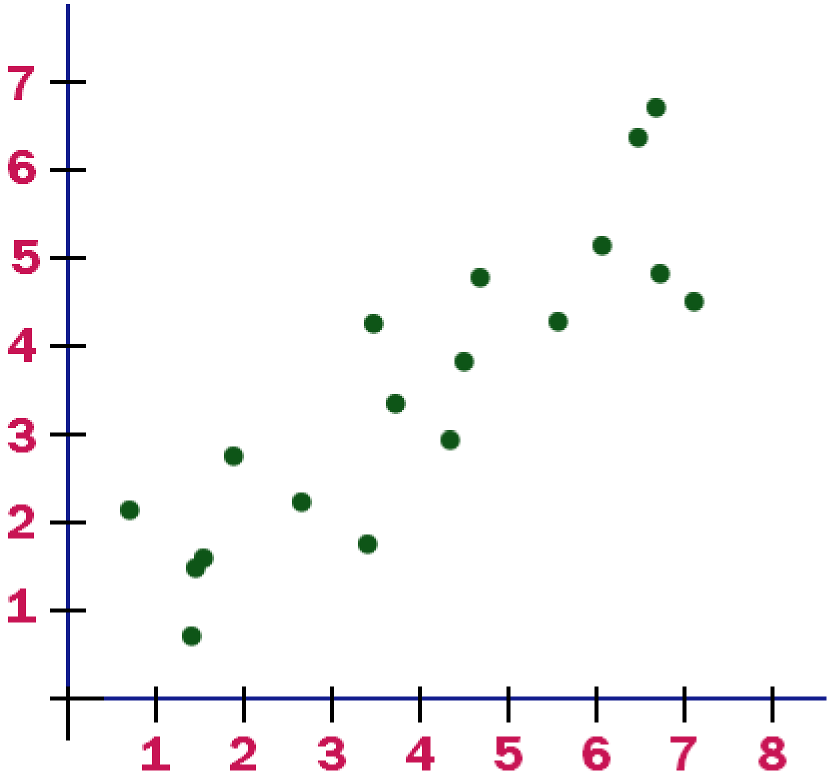
A strong positive correlation has variables that has the same changes, but the point are more close together and form a line.



Strong Positive Correlation

Weak Positive Correlation

A weak positive correlation has variables that has the same changes but the points on the graph are dispersed.



Weak Positive Correlation

Negative Correlation

In a negative correlation, as the values of one of the variables increase, the values of the second variable decrease or the value of one of the variables decreases, the value of the other variable increases. When the values of two variables x and y move in opposite direction, we say correlation is negative. That is in negative correlation, when there is an increase in x, there will be a decrease in y. Similarly when there is a decrease in x, there will be an increase in y increase.

Negative Correlation Example

When price increases, demand also decreases; when price decreases, demand also increases. So price and demand are negatively correlated.

Perfect Negative Correlation

The closer the correlation coefficient is either -1 or +1, the stronger the relationship is between the two variables. A perfect negative correlation of -1.0 indicated that for every member of the sample, higher score on one variable is related to a lower score on the other variable.

Correlation Data Sets

In statistics, some times we will have to study the relationship between two or more variables. The statistical technique used to study the relationships between the variables is called the correlation technique. Correlation analysis is the analysis of association between two or more variables. The tendency of two or more variables to vary together directly or inversely is called as correlation.

Two variables are said to be correlated, if the change in one of the variable results in a corresponding change in the other variable. That is, when two variables move together, they are said to be correlated.

Let us take an example to understand the term correlation. In a given data with heights and weights of students in a school, we can assume that students with a more height would have a more weight. Besides, it is assumed that students who have short height will have less weight.

Correlation Analysis

Correlation is a term that refers to the strength of a relationship between two variables. Correlation and regression analysis are related in the sense that both deal with relationships among variables. The correlation coefficient is a measure of linear association between two variables. Values of the correlation coefficient are always between -1 and +1. The value of -1 represents a perfect negative correlation while a value of +1 represents a perfect positive correlation. A value of 0 means that there is no relationship between the variables being tested.

Interpretation of coefficient of correlation based on the error likely

1. If the coefficient of correlation is less than the error likely, then its not significant
2. If the coefficient of correlation is more than six times the error likely, it is significant.
3. If the error is too small and coefficient of correlation is 0.5 or more then the coefficient of correlation is significant.

The values of r between 0 and 1 are said to have a limited degree of correlation. A limited degree of correlation may be positive or negative. Limited correlation can be high, moderate or low based on whether it is close to 1 or 0.

Covariance Correlation

Covariance and correlation are both describe the degree of similarity between two random variables. Suppose that X and Y are real-valued random variables for the experiment with means E(X), E(Y) and variances var(X), var(Y), respectively. The covariance of X and Y is defined by

cov(X, Y) = E[(X - E(X))(Y - E(Y))]

and the correlation of X and Y is defined by

cor(X, Y) = cov(X,Y)std(X)std(Y)cov(X,Y)std(X)std(Y).

Cross Correlation

The cross correlation function is a measure of the similarity between two data sets. One set is displaced related to the other, corresponding values of the two sets are multiplied together and the product are summed to give the value of the cross correlation. Whenever two sets are almost same, the product will be positive and the cross correlation is large. When set are unlike, some of the products will be positive and some negative and the sum will be small.

Correlation Examples
Given below are some examples to calculate correlation.

Solved Example

Question:

To determine the correlation value for the given set of X and Y values:

X Values	Y Values
21	2.5
23	3.1
37	4.2
19	5.6
24	6.4
33	8.4

Solution:

Let us count the number of values.
N = 6
Determine the values for XY, X^2, Y^2

X Value	Y Value	X*Y	X*X	Y*Y
21	2.5	52.5	441	6.25
23	3.1	71.3	529	9.61
37	4.2	155.4	1369	17.64
19	5.6	106.4	361	31.36
24	6.4	153.6	576	40.96
33	8.4	277.2	1089	70.56

Determine the following values $\sum X\sum X$, $\sum Y\sum Y$, $\sum XY\sum XY$, $\sum X_2\sum X2$, $\sum y_2\sum y2$.
$\sum X=157\sum X=157$
$\sum Y=30.2\sum Y=30.2$
$\sum XY=816.4\sum XY=816.4$

∑X2=4365∑X2=4365
∑Y2=176.38∑Y2=176.38

Correlation (r)

= N∑XY−(∑X)(∑Y)[N∑X2−(∑X)2][N∑Y2−(∑Y)2]/√N∑XY−(∑X)(∑Y)[N∑X2−(∑X)2][N∑Y2−(∑Y)2]

= 157(1541)(146.24)/√157(1541)(146.24)

(r)=0.33

Correlation Coefficient, *r* :

- The quantity *r*, called the *linear correlation coefficient*, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the *Pearson product moment correlation coefficient* in honor of its developer Karl Pearson.
- The mathematical formula for computing ***r*** is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right)-\left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right)-\left(\sum y\right)^2}}$$

where *n* is the number of pairs of data.

(Aren't you glad you have a graphing calculator that computes this formula?)

- The value of *r* is such that -1 ≤ *r* ≤ +1. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.
- *Positive correlation:* If *x* and *y* have a strong positive linear correlation, *r* is close to +1. An *r* value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between *x* and *y* variables such that as values

for *x* increases, values for *y* also increase.

- *Negative correlation:* If *x* and *y* have a strong negative linear correlation, *r* is close to -1. An *r* value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between *x* and *y* such that as values for *x* increase, values for *y* decrease.
- *No correlation:* If there is no linear correlation or a weak linear correlation, *r* is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables
- Note that *r* is a dimensionless quantity; that is, it does not depend on the units employed.
- A *perfect* correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If $r = +1$, the slope of this line is positive. If $r = -1$, the slope of this line is negative.
- A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Coefficient of Determination, r^2 or R^2:

- The *coefficient of determination*, r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.

- The *coefficient of determination* is the ratio of the explained variation to the total variation.
- The *coefficient of determination* is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between x and y.
- The *coefficient of determination* represents the percent of the data that is the closest to the line of best fit. For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.
- The *coefficient of determination* is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient: Pearson's correlation or Pearson correlation is a **correlation coefficient**commonly used in linear regression.

Sample question: Find the value of the correlation coefficient from the following table:

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87

6	59	81

Step 1:*Make a chart.* Use the given data, and add three more columns: xy, x^2, and y^2.

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99			
2	21	65			
3	25	79			
4	42	75			
5	57	87			
6	59	81			

Step 2::*Multiply x and y together to fill the xy column. For example, row 1 would be 43 × 99 = **4,257**.*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257		
2	21	65	1365		
3	25	79	1975		
4	42	75	3150		
5	57	87	4959		
6	59	81	4779		

Step 3: *Take the square of the numbers in the x column, and put the result in the x^2 column.*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	
2	21	65	1365	441	
3	25	79	1975	625	
4	42	75	3150	1764	
5	57	87	4959	3249	

6	59	81	4779	3481	

Step 4: *Take the square of the numbers in the y column, and put the result in the y^2 column.*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561

Step 5: *Add up all of the numbers in the columns and put the result at the bottom.^{2} column.* The Greek letter sigma (Σ) is a short way of saying “sum of.”

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

Step 6:*Use the following correlation coefficient formula.*

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

The answer is: **2868 / 5413.27 = 0.529809**

Click here if you want easy, step-by-step instructions for solving this formula.

From our table:

- $\Sigma x = 247$
- $\Sigma y = 486$
- $\Sigma xy = 20{,}485$
- $\Sigma x^2 = 11{,}409$
- $\Sigma y^2 = 40{,}022$
- n is the sample size, in our case = 6

The correlation coefficient =

- $6(20{,}485) – (247 \times 486) / [\sqrt{[[6(11{,}409) – (247^2)] \times [6(40{,}022) – 486^2]]}]$

=0.5298

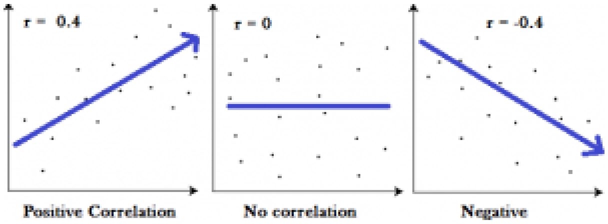
The range of the correlation coefficient is from -1 to 1. Our result is 0.5298 or 52.98%, which means the variables have a moderate positive correlation.

Pearson’s Correlation Coefficient returns a value of between -1 and +1. A -1 means there is a strong negative correlation and +1 means that there is a strong positive correlation. This can initially be a little hard to wrap your head around (who likes to deal with negative numbers?). The Political Science Department at Quinnipiac University posted this useful list of the meaning of Pearson’s Correlation coefficients. They note that these are “**crude estimates**” for interpreting strengths of correlations using Pearson’s Correlation:

r value =	
+.70 or higher	Very strong positive relationship
+.40 to +.69	Strong positive relationship
+.30 to +.39	Moderate positive relationship
+.20 to +.29	weak positive relationship
+.01 to +.19	No or negligible relationship
0	No relationship
-.01 to -.19	No or negligible relationship
-.20 to -.29	weak negative relationship
-.30 to -.39	Moderate negative relationship
-.40 to -.69	Strong negative relationship

-.70 or higher	Very strong negative relationship

It may be helpful to see graphically what these correlations look like:



Graphs showing a correlation of -1 (a negative correlation), 0 and +1 (a positive correlation)

The images show that a strong negative correlation means that the graph has a downward slope from left to right: as the x-values increase, the y-values get smaller. A strong positive correlation means that the graph has an upward slope from left to right: as the x-values increase, the y-values get larger.

Regression Definition:

A regression is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables.

Regression Formula:

Regression Equation(y) = a + bx Slope(b) = (NΣXY - (ΣX)(ΣY)) / (NΣX^2 - (ΣX)^2) Intercept(a) = (ΣY - b(ΣX)) / N **Where,**

x and y are the variables. b = The slope of the regression line a = The intercept point of the regression line and the y axis. N = Number of values or elements X = First Score Y = Second Score ΣXY = Sum of the product of first and Second Scores ΣX = Sum of First Scores ΣY = Sum of Second Scores ΣX^2 = Sum of square First Scores

Regression Example:

To find the Simple/Linear Regression of

X Values	Y Values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

To find regression equation, we will first find slope, intercept and use it to form regression equation.

Step 1:

Count the number of values. N = 5

Step 2:

Find XY, X^2 See the below table

X Value	Y Value	X*Y	X*X
60	3.1	60 * 3.1 = 186	60 * 60 = 3600
61	3.6	61 * 3.6 = 219.6	61 * 61 = 3721
62	3.8	62 * 3.8 = 235.6	62 * 62 = 3844
63	4	63 * 4 = 252	63 * 63 = 3969
65	4.1	65 * 4.1 = 266.5	65 * 65 = 4225

Step 3:

Find ΣX, ΣY, ΣXY, ΣX^2. ΣX = 311 ΣY = 18.6 ΣXY = 1159.7 ΣX^2 = 19359

Step 4:

Substitute in the above slope formula given. Slope(b) = (NΣXY - (ΣX)(ΣY)) / (NΣX^2 - (ΣX)^2) = ((5)*(1159.7)-(311)*(18.6))/((5)*(19359)-(311)^2) = (5798.5 - 5784.6)/(96795 - 96721) = 13.9/74 = 0.19

Step 5:

Now, again substitute in the above intercept formula given. Intercept(a) = (ΣY - b(ΣX)) / N = (18.6 - 0.19(311))/5 = (18.6 - 59.09)/5 = -40.49/5 = -8.098

Step 6:

Then substitute these values in regression equation formula Regression Equation(y) = a + bx = -8.098 + 0.19x. Suppose if we want to know the approximate y value for the variable x = 64. Then we can substitute the value in the above equation. Regression Equation(y) = a + bx = -8.098 + 0.19(64). = -8.098 + 12.16 = 4.06 This example will guide you to find the relationship between two variables by calculating the Regression from the above steps.

UNIT IV

A **parameter** is a characteristic of a**population**. A **statistic** is a characteristic of a**sample**. Inferential **statistics** enables you to make an educated guess about a **population parameter** based on a **statistic** computed from a **sample** randomly drawn from that **population**

Tests of Significance

Once sample data has been gathered through an observational study or experiment, statistical inference allows analysts to assess evidence in favor or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as ***tests of significance***.

Every test of significance begins with a ***null hypothesis*** $\boldsymbol{H_0}$. H_0 represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug. We would write H_0: there is no difference between the two drugs on average.

The ***alternative hypothesis***, H_a, is a statement of what a statistical hypothesis test is set up to establish. For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect, on average, compared to that of the current drug. We would write H_a: the two drugs have different effects, on average. The alternative hypothesis might also be that the new drug is better, on average, than the current drug. In this case we would write H_a: the new drug is better than the current drug, on average.

The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either "reject H_0 in favor of H_a" or "do not reject H_0"; we never conclude "reject H_a", or even "accept H_a".

If we conclude "do not reject H_0", this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against H_0 in favor of H_a; rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.

Hypotheses are always stated in terms of population parameter, such as the mean μ . An alternative hypothesis may be ***one-sided*** or ***two-sided***. A one-sided hypothesis claims that a parameter is either larger *or* smaller than the value given by the null hypothesis. A two-sided hypothesis claims that a parameter is simply *not equal* to the value given by the null hypothesis -- the direction does not matter.

Hypotheses for a one-sided test for a population mean take the following form:
H_0: μ = k
H_a: μ > k
or

H_0: μ = k
H_a: μ < k.

Hypotheses for a two-sided test for a population mean take the following form:
H_0: μ = k
H_a: $\mu \neq$ k.

A ***confidence interval*** gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data.

Example

Suppose a test has been given to all high school students in a certain state. The mean test score for the entire state is 70, with standard deviation equal to 10. Members of the school board suspect that female students have a higher mean score on the test than male students, because the mean score $\bar{x}$ from a random sample of 64 female students is equal to 73. Does this provide strong evidence that the overall mean for female students is higher?

The null hypothesis H_0 claims that there is no difference between the mean score for female students and the mean for the entire population, so that μ = 70. The alternative hypothesis claims that the mean for female students is higher than the entire student population mean, so that μ > 70.

Significance Levels

The ***significance level*** α for a given hypothesis test is a value for which a *P-value* less than or equal to α is considered statistically significant. Typical values for α are 0.1, 0.05, and 0.01. These values correspond to the probability of observing such an extreme value by chance. In the test score example above, the *P-value*is 0.0082, so the probability of observing such a value by chance is less that 0.01, and the result is significant at the 0.01 level.

In a one-sided test, α corresponds to the critical value z^* such that $P(Z \geq z^*) = \alpha$. For example, if the desired significance level for a result is 0.05, the corresponding value for z must be greater than or equal to $z^* = 1.645$ (or less than or equal to -1.645 for a one-sided alternative claiming that the mean is less than the null hypothesis). For a two-sided test, we are interested in the probability that $2P(Z \geq z^*) = \alpha$, so the critical value z^* corresponds to the $\alpha/2$ significance level. To achieve a significance level of 0.05 for a two-sided test, the absolute value of the test statistic (|z|) must be greater than or equal to the critical value 1.96 (which corresponds to the level 0.025 for a one-sided test).

Another interpretation of the significance level α, based in ***decision theory***, is that α corresponds to the value for which one chooses to reject or accept the null hypothesis H_0. In the above example, the value 0.0082 would result in rejection of the null hypothesis at the 0.01 level. The probability that this is a mistake -- that, in fact, the null hypothesis is true given the z-statistic -- is less than 0.01. In decision theory, this is known as a ***Type I error***. The probability of a Type I error is equal to the significance level α, and the probability of rejecting the null hypothesis when it is in fact false (a correct decision) is equal to 1 - α. To minimize the probability of Type I error, the significance level is generally chosen to be small.

Example

Of all of the individuals who develop a certain rash, suppose the mean recovery time for individuals who do not use any form of treatment is 30 days with standard deviation equal to 8. A pharmaceutical company manufacturing a certain cream wishes to determine whether the cream shortens, extends, or has no effect on the recovery time. The company chooses a random sample of 100 individuals who have used the cream, and determines that the mean recovery time for these individuals was 28.5 days. Does the cream have any effect?

Since the pharmaceutical company is interested in *any* difference from the mean recovery time for all individuals, the alternative hypothesis H_a is two-sided: $\mu \neq 30$. The test statistic is calculated to be z = (28.5 - 30)/(8/sqrt(100)) = -1.5/0.8 = -1.875. The *P-value* for this statistic is $2P(Z \geq 1.875)$ = 2(1 - P((Z < 1.875) = 2(1- 0.9693) = 2(0.0307) = 0.0614. This is not significant at the 0.05 level, although it is significant at the 0.1 level.

Decision theory is also concerned with a second error possible in significance testing, known as ***Type II error***. Contrary to Type I error, Type II error is the error made when the null hypothesis is incorrectly accepted. The probability of correctly rejecting the null hypothesis when it is false, the complement of the Type II error, is known as the ***power*** of a test. **Formally defined, the *power* of**

a test is the probability that a fixed level α significance test will reject the null hypothesis H_0 when a particular alternative value of the parameter is true.

Hypothesis Testing Examples (One Sample Z Test)

The one sample z test isn't used very often (because we rarely know the actual population standard deviation). However, it's a good idea to understand how it works as it's one of the simplest tests you can perform in hypothesis testing. In English class you got to learn the basics (like grammar and spelling) before you could write a story; think of one sample z tests as the foundation for understanding more complex hypothesis testing. This page contains two hypothesis testing examples for one sample z-tests.

A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with a standard deviation of 15.

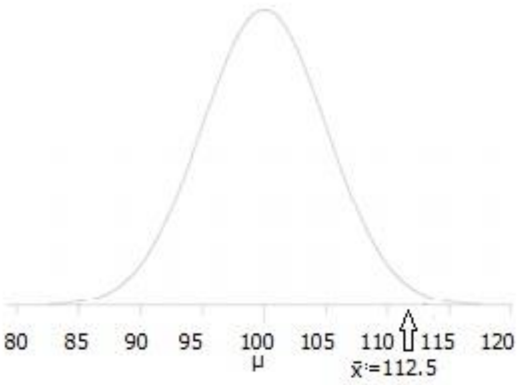
Step 1: State the Null hypothesis. The accepted fact is that the population mean is 100, so: H_0: $\mu=100$.

Step 2: State the Alternate Hypothesis. The claim is that the students have above average IQ scores, so:

H_1: $\mu > 100$.

The fact that we are looking for scores "greater than" a certain point means that this is a one-tailed test.

Step 3: Draw a picture to help you visualize the problem.



Step 4: State the alpha level. If you aren't given an alpha level, use 5% (0.05).

Step 5: Find the rejection region area (given by your alpha level above) from the z-table. An area of .05 is equal to a z-scoreof 1.645.

$$Z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

Step 6: Find the test statistic using this formula:

For this set of data: z= (112.5-100) / (15/√30)=4.56.

Step 6: If Step 6 is greater than Step 4, reject the null hypothesis. If it's less than Step 4, you cannot reject the null hypothesis. In this case, it is greater, so you can reject the null.

Hypothesis Testing of the Difference Between Two Population Means

This file is part of a program based on the Bio 4835 Biostatistics class taught at Kean University in Union, New Jersey. The course uses the following text:
Daniel, W. W. 1999. Biostatistics: a foundation for analysis in the health sciences. New York: John Wiley and Sons.
The file follows this text very closely and readers are encouraged to consult the text for further information.

B) Hypothesis testing of the difference between two population means

This is a two sample z test which is used to determine if two population means are equal or unequal. There are three possibilities for formulating hypotheses.

l. $H_0 : \mu_1 = \mu_2$ $\quad H_A : \mu_1 \neq \mu_2$

2. $H_0 : \mu_1 \geq \mu_2$ $\quad H_A : \mu_1 < \mu_2$

3. $H_0 : \mu_1 \leq \mu_2$ $\quad H_A : \mu_1 > \mu_2$

Procedure

The same procedure is used in three different situations

- Sampling is from normally distributed populations with known variances

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

- Sampling from normally distributed populations where population variances are unknown
    - population variances equal

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$

This is with *t* distributed as Student's *t* distribution with (n_1 + n_2 - 2) degrees of freedom and a pooled variance.

- population variances unequal

When population variances are unequal, a distribution of *t'* is used in a manner similar to calculations of confidence intervals in similar circumstances.

- Sampling from populations that are not normally distributed

If both sample sizes are 30 or larger the central limit theorem is in effect. The test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

If the population variances are unknown, the sample variances are used.

Sampling from normally distributed populations with population variances known

Example 7.3.1

Serum uric acid levels

Is there a difference between the means between individuals with Down's syndrome and normal individuals?

(1) Data

$\bar{x}_1$= 4.5 n_1 = 12 σ_1^2 = 1
$\bar{x}_2$ = 3.4 n_2 = 15 σ_2^2 = 1.5
α = .05

(2) Assumptions

- two independent random samples
- each drawn from a normally distributed population

(3) Hypotheses

H_0 : $\mu_1 = \mu_2$

H_A : $\mu_1 \neq \mu_2$

(4) Test statistic

This is a two sample z test.

(a) Distribution of test statistic

If the assumptions are correct and H_0 is true, the test statistic is distributed as the normal distribution.

(b) Decision rule

With α = .05, the critical values of z are -1.96 and +1.96. We reject H_0 if z < -1.96 or z > +1.96.

(5) Calculation of test statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

$$z = \frac{(4.5 - 3.4) - 0}{\sqrt{1/12 + 1.5/15}} = \frac{1.1}{.4282} = 2.57$$

(6) Statistical decision

Reject H_0 because 2.57 > 1.96.

(7) Conclusion

From these data, it can be concluded that the population means are not equal. A 95% confidence interval would give the same conclusion.

p = .0102.

Sampling from normally distributed populations with unknown variances

With equal population variances, we can obtain a pooled value from the sample variances.

Example 7.3.2

Lung destructive index

We wish to know if we may conclude, at the 95% confidence level, that smokers, in general, have greater lung damage than do non-smokers.

(1) Data

Smokers: $\bar{x}_1 = 17.5$ $n_1 = 16$ $s_1^2 = 4.4752$

Non-Smokers: $\bar{x}_2 = 12.4$ $n_2 = 9$ $s_2^2 = 4.8492$

$\alpha = .05$

Calculation of Pooled Variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_p^2 = \frac{(15)(4.4711) + (8)(4.8492)}{16 + 9 - 2}$$

$$s_p^2 = \frac{299.86 + 188.12}{23}$$

$$s_p^2 = 21.2165$$

(2) Assumptions

- independent random samples
- normal distribution of the populations
- population variances are equal

(3) Hypotheses

$$H_0 : \mu_1 \le \mu_2$$

$H_A : \mu_1 > \mu_2$

(4) Test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2 / n_1 + s_p^2 / n_2}}$$

(a) Distribution of test statistic

If the assumptions are met and H_0 is true, the test statistic is distributed as Student's t distribution with 23 degrees of freedom.

(b) Decision rule

With α = .05 and df = 23, the critical value of *t* is 1.7139. We reject H_0 if t > 1.7139.

(5) Calculation of test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2 / n_1 + s_p^2 / n_2}}$$

$$t = \frac{(17.5 - 12.4) - 0}{\sqrt{21.2165/16 + 21.2165/9}} = \frac{5.1}{1.92} = 2.6563$$

(6) Statistical decision

Reject H_0 because 2.6563 > 1.7139.

(7) Conclusion

On the basis of the data, we conclude that $\mu_1 > \mu_2$.

Actual values
t = 2.6558
p = .014

Sampling from populations that are not normally distributed

Example 7.3.4

These data were obtained in a study comparing persons with disabilities with persons without disabilities. A scale known as the Barriers to Health Promotion Activities for Disabled Persons (BHADP) Scale gave the data. We wish to know if we may conclude, at the 99% confidence level, that persons with disabilities score higher than persons without disabilities.

(1) Data

Disabled: $\bar{x}_1 = 31.83$ $n_1 = 132$ $s_1 = 7.93$
Nondisabled: $\bar{x}_2 = 25.07$ $n_2 = 137$ $s_2 = 4.80$
$\alpha = .01$

(2) Assumptions

- independent random samples

(3) Hypotheses

$H_0 : \mu_1 \leq \mu_2$
$H_A : \mu_1 > \mu_2$

(4) Test statistic

Because of the large samples, the central limit theorem permits calculation of the z score as opposed to using *t*. The z score is calculated using the given sample standard deviations.

(a) Distribution of test statistic

If the assumptions are correct and H_0 is true, the test statistic is approximately normally distributed

(b) Decision rule

With $\alpha = .01$ and a one tail test, the critical value of z is 2.33. We reject H_0 $z > 2.33$.

(5) Calculation of test statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}}$$

$$z = \frac{(31.83 - 25.07) - 0}{\sqrt{(7.93)^2 / 132 + (4.80)^2 / 137}} = \frac{6.76}{.8029} = 8.42$$

(6) Statistical decision

Reject H_0 because 8.42 > 2.33.

(7) Conclusion

On the basis of these data, the average persons with disabilities score higher on the BHADP test than do the nondisabled persons.

Actual values
z = 8.42
p = 1.91 x 10-17

Paired comparisons

Sometimes data comes from nonindependent samples. An example might be testing "before and after" of cosmetics or consumer products. We could use a single random sample and do "before and after" tests on each person. A hypothesis test based on these data would be called a *paired comparisons test*. Since the observations come in pairs, we can study the difference, d, between the samples. The difference between each pair of measurements is called di.

Test statistic

With a population of n pairs of measurements, forming a simple random sample from a normally distributed population, the mean of the difference, μ_d , is tested using the following implementation of *t*.

$$t = \frac{\bar{d} - \mu_{d_0}}{s_{\bar{d}}}$$

$\bar{d}$ is the sample mean difference

μ_{d_0} is the hypothesized mean difference

$s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$ -- the standard error

n is the number of sample differences

s_d is the standard deviation of the sample diffences

Paired comparisons

Example 7.4.1

Very-low-calorie diet (VLCD) Treatment

Table gives B (before) and A (after) treatment data for obese female patients in a weight-loss program.

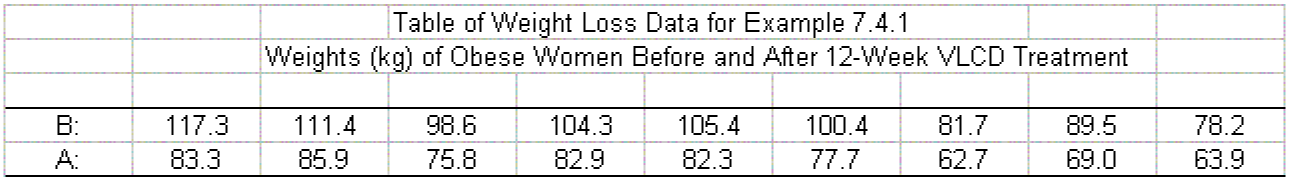


			Table of Weight Loss Data for Example 7.4.1						
		Weights (kg) of Obese Women Before and After 12-Week VLCD Treatment							
B:	117.3	111.4	98.6	104.3	105.4	100.4	81.7	89.5	78.2
A:	83.3	85.9	75.8	82.9	82.3	77.7	62.7	69.0	63.9

We calculate di = A-B for each pair of data resulting in negative values meaning that the participants lost weight.

We wish to know if we may conclude, at the 95% confidence level, that the treatment is effective in causing weight reduction in these people.

(1) Data

Values of di are calculated by subtracting each A from each B to give a negative number. On the TI-83 calculator place the A data in L1 and the B data in L2. Then make L3 = L1 - L2 and the calculator does each calculation automatically.

In Microsoft Excel put the A data in column A and the B data in column B, without using column headings so that the first pair of data are on line 1. In cell C1, enter the following formula: =a1-b1. This calculates the difference, di, for B - A. Then copy the formula down column C until the rest of the differences are calculated.

n = 9
α = .05

(2) Assumptions

- the observed differences are a simple random sample from a normally distributed population of differences

(3) Hypotheses

H_0: $\mu_d \geq 0$
H_A: $\mu_d < 0$ (meaning that the patients lost weight)

(4) Test statistic

The test statistic is *t* which is calculated as

$$t = \frac{\bar{d} - \mu_{d_0}}{s_{\bar{d}}}$$

(a) Distribution of test statistic

The test statistic is distributed as Student's t with 8 degrees of freedom

(b) Decision rule

With α = .05 and 8 df the critical value of t is -1.8595. We reject H_0 if t < -1.8595.

(5) Calculation of test statistic

$$\bar{d} = \frac{\sum di}{n} = \frac{-203.3}{9} = -22.5889$$

$$s_{\bar{d}}^2 = 28.2961$$

$$t = \frac{\bar{d} - \mu_{d_0}}{s_{\bar{d}}} = \frac{-22.5899 - 0}{\sqrt{28.2961/9}} = -12.7395$$

(6) Statistical decision

Reject H_0 because -12.7395 < -1.8595
p = 6.79 x 10-7

(7) Conclusion

On the basis of these data, we conclude that the diet program is effective.

UNIT V

The Chi Square Statistic

Types of Data:

There are basically two types of random variables and they yield two types of data: numerical and categorical. A chi square (X^2) statistic is used to investigate whether distributions of categorical variables differ from one another. Basically categorical variable yield data in the categories and numerical variables yield data in numerical form. Responses to such questions as "What is your major?" or Do you own a car?" are categorical because they yield data such as "biology" or "no." In contrast, responses to such questions as "How tall are you?" or "What is your G.P.A.?" are numerical. Numerical data can be either discrete or continuous. The table below may help you see the differences between these two variables.

Data Type	**Question Type**	**Possible Responses**
Categorical	What is your sex?	male or female
Numerical	Disrete- How many cars do you own?	two or three

Numerical	Continuous - How tall are you?	72 inches

Notice that discrete data arise fom a counting process, while continuous data arise from a measuring process.

The Chi Square statistic compares the tallies or counts of categorical responses between two (or more) independent groups. (note: Chi square tests can only be used on actual numbers and not on percentages, proportions, means, etc.)

2 x 2 Contingency Table

There are several types of chi square tests depending on the way the data was collected and the hypothesis being tested. We'll begin with the simplest case: a 2 x 2 contingency table. If we set the 2 x 2 table to the general notation shown below in Table 1, using the letters a, b, c, and d to denote the contents of the cells, then we would have the following table:

Table 1. General notation for a 2 x 2 contingency table.
Variable 1

Variable 2	Data type 1	Data type 2	Totals
Category 1	a	b	a + b
Category 2	c	d	c + d
Total	a + c	b + d	a + b + c + d = N

For a 2 x 2 contingency table the Chi Square statistic is calculated by the formula:

$$\chi^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)}$$

Note: notice that the four components of the denominator are the four totals from the table columns and rows.

Suppose you conducted a drug trial on a group of animals and you hypothesized that the animals receiving the drug would show increased heart rates compared to those that did not receive the drug. You conduct the study and collect the following data:

Ho: The proportion of animals whose heart rate increased is independent of drug treatment.

Ha: The proportion of animals whose heart rate increased is associated with drug treatment.

Table 2. Hypothetical drug trial results.

	Heart Rate Increased	No Heart Rate Increase	Total
Treated	36	14	50
Not treated	30	25	55
Total	66	39	105

Applying the formula above we get:

Chi square = 105[(36)(25) - (14)(30)]^2 / (50)(55)(39)(66) = 3.418

Before we can proceed we eed to know how many degrees of freedom we have. When a comparison is made between one sample and another, a simple rule is that the degrees of freedom equal (number of columns minus one) x (number of rows minus one) not counting the totals for rows or columns. For our data this gives (2-1) x (2-1) = 1.

We now have our chi square statistic (x^2 = 3.418), our predetermined alpha level of significance (0.05), and our degrees of freedom (df = 1). Entering the Chi square distribution table with 1 degree of freedom and reading along the row we find our value of x^2 (3.418) lies between 2.706 and 3.841. The corresponding probability is between the 0.10 and 0.05 probability levels. That means that the p-value is above 0.05 (it is actually 0.065). Since a p-value of 0.65 is greater than the conventionally accepted significance level of 0.05 (i.e. $p > 0.05$) we fail to reject the null hypothesis. In other words, there is no statistically significant difference in the proportion of animals whose heart rate increased.

What would happen if the number of control animals whose heart rate increased dropped to 29 instead of 30 and, consequently, the number of controls whose hear rate did not increase changed from 25 to 26? Try it. Notice that the new x^2 value is 4.125 and this value exceeds the table value of 3.841 (at 1 degree of freedom and an alpha level of 0.05). This means that $p < 0.05$ (it is now0.04) and we reject the null hypothesis in favor of the alternative hypothesis - the heart rate of animals is different between the treatment groups. When $p < 0.05$ we generally refer to this as a significant difference.

Table 3. Chi Square distribution table.

probability level (alpha)

Df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

To make the chi square calculations a bit easier, plug your observed and expected values into the following applet. Click on the cell and then enter the value. Click the compute button on the lower right corner to see the chi square value printed in the lower left hand coner.

Chi Square Goodness of Fit (One Sample Test)

This test allows us to compae a collection of categorical data with some theoretical expected distribution. This test is often used in genetics to compare the results of a cross with the theoretical distribution based on genetic theory. Suppose you preformed a simpe monohybrid cross between two individuals that were heterozygous for the trait of interest.

Aa x Aa

The results of your cross are shown in Table 4.

Table 4. Results of a monohybrid coss between two heterozygotes for the 'a' gene.

	A	a	Totals
A	10	42	52
a	33	15	48
Totals	43	57	100

The penotypic ratio 85 of the A type and 15 of the a-type (homozygous recessive). In a monohybrid cross between two heterozygotes, however, we would have predicted a 3:1 ratio of phenotypes. In other words, we would have expected to get 75 A-type and 25 a-type. Are or resuls different?

$$\chi^2 = \sum \frac{(observed - \exp ected)^2}{\exp ected}$$

Calculate the chi square statistic x^2 by completing the following steps:

1. For each *observed* number in the table subtract the corresponding *expected* number (*O* — *E*).
2. Square the difference [$(O — E)^2$].
3. Divide the squares obtained for each cell in the table by the *expected* number for that cell [$(O - E)^2 / E$].
4. Sum all the values for $(O - E)^2 / E$. This is the chi square statistic.

For our example, the calculation would be:

	Observed	Expected	(O — E)	$(O — E)^2$	$(O — E)^2/ E$
A-type	85	75	10	100	1.33
a-type	15	25	10	100	4.0
Total	100	100			5.33

$x^2 = 5.33$

We now have our chi square statistic ($x^2 = 5.33$), our predetermined alpha level of significalnce (0.05), and our degrees of freedom (df =1). Entering the Chi square distribution table with 1 degree of freedom and reading along the row we find our value of x^2 5.33) lies between 3.841 and 5.412. The corresponding probability is 0.05<P<0.02. This is smaller than the conventionally accepted significance level of 0.05 or 5%, so the null hypothesis that the two distributions are the same is rejected. In other words, when the computed x^2 statistic exceeds the critical value in the table for a 0.05 probability level, then we can reject the null hypothesis of equal distributions. Since our x^2 statistic (5.33) exceeded the critical value for 0.05 probability level (3.841) we can reject the null hypothesis that the observed values of our cross are the same as the theoretical distribution of a 3:1 ratio.

Table 3. Chi Square distribution table.

probability level (alpha)

Df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

To put this into context, it means that we do not have a 3:1 ratio of A_ to aa offspring.

To make the chi square calculations a bit easier, plug your observed and expected values into the following java applet.

Click on the cell and then enter the value. Click the compute button on the lower right corner to see the chi square value printed in the lower left hand coner.

Chi Square Test of Independence

For a contingency table that has *r* rows and *c* columns, the chi square test can be thought of as a test of independence. In a test ofindependence the null and alternative hypotheses are:

Ho: The two categorical variables are independent.

Ha: The two categorical variables are related.

We can use the equation Chi Square = the sum of all the□$(f_o - f_e)^2 / f_e$

Here f_o denotes the frequency of the observed data and f_e is the frequency of the expected values. The general table would look something like the one below:

	Category I	Category II	Category III	Row Totals
Sample A	a	b	c	a+b+c
Sample B	d	e	f	d+e+f
Sample C	g	h	i	g+h+i
Column Totals	a+d+g	b+e+h	c+f+i	a+b+c+d+e+f+g+h+i=N

Now we need to calculate the expected values for each cell in the table and we can do that using the the row total times the column total divided by the grand total (N). For example, for cell a the expected value would be (a+b+c)(a+d+g)/N.

Once the expected values have been calculated for each cell, we can use the same procedure are before for a simple 2 x 2 table.

Observed	Expected	|O - E|	$(O — E)^2$	$(O — E)^2/ E$

Suppose you have the following categorical data set.

Table . Incidence of three types of malaria in three tropical regions.

	Asia	Africa	South America	Totals
Malaria A	31	14	45	90
Malaria B	2	5	53	60
Malaria C	53	45	2	100
Totals	86	64	100	250

We could now set up the following table:

Observed	Expected	|O -E|	(O — E)^2	(O — E)^2/ E
31	30.96	0.04	0.0016	0.0000516
14	23.04	9.04	81.72	3.546
45	36.00	9.00	81.00	2.25
2	20.64	18.64	347.45	16.83
5	15.36	10.36	107.33	6.99
53	24.00	29.00	841.00	35.04
53	34.40	18.60	345.96	10.06
45	25.60	19.40	376.36	14.70
2	40.00	38.00	1444.00	36.10

Chi Square = 125.516

Degrees of Freedom = (c - 1)(r - 1) = 2(2) = 4

Table 3. Chi Square distribution table.

probability level (alpha)

Df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

Reject Ho because 125.516 is greater than 9.488 (for alpha = 0.05)

Thus, we would reject the null hypothesis that there is no relationship between location and type of malaria. Our data tell us there is a relationship between type of malaria and location, but that's all it says.

One-Way Analysis of Variance (ANOVA) Example Problem

Introduction

Analysis of Variance (ANOVA) is a hypothesis-testing technique used to test the equality of two or more population (or treatment) means by examining the variances of samples that are taken. ANOVA allows one to determine whether the differences between the samples are simply due to random error (sampling errors) or whether there are systematic treatment effects that causes the mean in one group to differ from the mean in another.

Most of the time ANOVA is used to compare the equality of three or more means, however when the means from two samples are compared using ANOVA it is equivalent to using a t-test to compare the means of independent samples.

ANOVA is based on comparing the variance (or variation) *between* the data samples to variation *within* each particular sample. If the between variation is much larger than the within variation, the means of different samples will not be equal. If the between and within variations are approximately the same size, then there will be no significant difference between sample means.

Assumptions of ANOVA:
(i) All populations involved follow a normal distribution.
(ii) All populations have the same variance (or standard deviation).
(iii) The samples are randomly selected and independent of one another.

Since ANOVA assumes the populations involved follow a normal distribution, ANOVA falls into a category of hypothesis tests known as parametric tests. If the populations involved did not follow a normal distribution, an ANOVA test could not be used to examine the equality of the sample means. Instead, one would have to use a non-parametric test (or distribution-free test), which is a more general form of hypothesis testing that does not rely on distributional assumptions.

Example

Consider this example:
Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use α = 5%.

Table ANOVA.1

	Compact cars	Midsize cars	Full-size cars
	643	469	484
	655	427	456
	702	525	402
$\overline{X}$	666.67	473.67	447.33
S	31.18	49.17	41.68

(1.) State the null and alternative hypotheses
The null hypothesis for an ANOVA always assumes the population means are equal. Hence, we may write the null hypothesis as:

H_0: $\mu_1 = \mu_2 = \mu_3$ - The mean head pressure is statistically equal across the three types of cars.

Since the null hypothesis assumes all the means are equal, we could reject the null hypothesis if only mean is not equal. Thus, the alternative hypothesis is:
H_a: At least one mean pressure is not statistically equal.

(2.) Calculate the appropriate test statistic
The test statistic in ANOVA is the ratio of the *between* and *within* variation in the data. It follows an F distribution.

Total Sum of Squares – the total variation in the data. It is the sum of the between and within variation.

Total Sum of Squares (SST) = $\sum_{i=1}^{r}\sum_{j=1}^{c}(X_{ij} - \overline{\overline{X}})^2$, where r is the number of rows in the table, c is the number of columns, $\overline{\overline{X}}$ is the grand mean, and X_{ij} is the i th observation in the j th column.

Using the data in Table ANOVA.1 we may find the grand mean:

$$\overline{\overline{X}} = \frac{\sum X_{ij}}{N} = \frac{(643+655+702+469+427+525+484+456+402)}{9} = 529.22$$

SST =

$$(643-529.22)^2 + (655-529.22)^2 + (702-529.22)^2 + (469-529.22)^2 + \ldots + (402-529.22)^2 = 96303.55$$

Between Sum of Squares (or Treatment Sum of Squares) – variation in the data between the different samples (or treatments).

Treatment Sum of Squares (SSTR) = $\sum r_j(\overline{X}_j - \overline{\overline{X}})^2$, where r_j is the number of rows in the j th treatment and $\overline{X}_j$ is the mean of the j th treatment.

Using the data in Table ANOVA.1,

SSTR = $[3*(666.67-529.22)^2]+[3*(473.67-529.22)^2]+[3*(447.33-529.22)^2] = 86049.55$

Within variation (or Error Sum of Squares) – variation in the data from each individual treatment.

Error Sum of Squares (SSE) = $\sum\sum(X_{ij} - \overline{X}_j)^2$

From Table ANOVA.1,

SSE=$[(643-666.67)^2+(655-666.67)^2+(702-666.67)^2]+$
$[(469-473.67)^2+(427-473.67)^2+(525-473.67)^2]+$
$[(484-447.33)^2+(456-447.33)^2+(402-447.33)^2]$ = 10254.

Note that SST = SSTR + SSE (96303.55 = 86049.55 + 10254).

Hence, you only need to compute any two of three sources of variation to conduct an ANOVA. Especially for the first few problems you work out, you should calculate all three for practice.

The next step in an ANOVA is to compute the "average" sources of variation in the data using SST, SSTR, and SSE.

Total Mean Squares (MST) $= \dfrac{SST}{N-1}$ → "average total variation in the data" (N is the total number of observations)

MST = $\dfrac{96303.55}{(9-1)} = 12037.94$

Mean Square Treatment (MSTR) = $\dfrac{SSTR}{c-1}$ → "average between variation" (c is the number of columns in the data table)

MSTR = $\dfrac{86049.55}{(3-1)}$ = 43024.78

Mean Square Error (MSE) = $\dfrac{SSE}{N-c}$ → "average within variation"

MSE = $\dfrac{10254}{(9-3)} = 1709$

Note: MST $\neq$ MSTR + MSE

The test statistic may now be calculated. For a one-way ANOVA the test statistic is equal to the ratio of MSTR and MSE. This is the ratio of the "average between variation" to the "average within variation." In addition, this ratio is known to follow an F distribution. Hence,

F = $\frac{MSTR}{MSE} = \frac{43024.78}{1709} = 25.17$. The intuition here is relatively straightforward. If the average between variation rises relative to the average within variation, the F statistic will rise and so will our chance of rejecting the null hypothesis.

(3.) Obtain the Critical Value
To find the critical value from an F distribution you must know the numerator (MSTR) and denominator (MSE) degrees of freedom, along with the significance level.

F^{CV} has df1 and df2 degrees of freedom, where df1 is the numerator degrees of freedom equal to c-1 and df2 is the denominator degrees of freedom equal to N-c.

In our example, df1 = 3 - 1 = 2 and df2 = 9 - 3 = 6. Hence we need to find $F_{2,6}^{CV}$ corresponding to α = 5%. Using the F tables in your text we determine that $F_{2,6}^{CV} =$ 5.14.

(4.) Decision Rule
You reject the null hypothesis if: F (observed value) > F^{CV} (critical value). In our example 25.17 > 5.14, so we reject the null hypothesis.

(5.) Interpretation
Since we rejected the null hypothesis, we are 95% confident (1-α) that the mean head pressure is not statistically equal for compact, midsize, and full size cars. However, since only one mean must be different to reject the null, we do not yet know which mean(s) is/are different. In short, an ANOVA test will test us that *at least one mean is different*, but an additional test must be conducted to determine which mean(s) is/are different.

Determining Which Mean(s) Is/Are Different

If you fail to reject the null hypothesis in an ANOVA then you are done. You know, with some level of confidence, that the treatment means are statistically equal. However, if you reject the null then you must conduct a separate test to determine which mean(s) is/are different.

There are several techniques for testing the differences between means, but the most common test is the Least Significant Difference Test.

Least Significant Difference (LSD) for a *balanced* sample: $\sqrt{\frac{2*MSE*F_{1,N-c}}{r}}$, where MSE is the mean square error and r is the number of rows in each treatment.

In the example above, LSD = $\sqrt{\frac{(2)(1709)(5.99)}{3}}$ = 82.61

Thus, if the absolute value of the difference between any two treatment means is greater than 82.61, we may conclude that they are not statistically equal.

Compact cars vs. Midsize cars:
$|666.67 - 473.67|$ = 193. Since 193 > 82.61 → mean head pressure is statistically different between compact and midsize cars.

Midsize cars vs. Full-size cars:
$|473.67 - 447.33|$ = 26.34. Since 26.34 < 82.61 → mean head pressure is statistically equal between midsize and full-size cars.

Compact vs. Full-size:
Work this on your own.

One-way ANOVA in Excel

You may conduct a one-way ANOVA using Excel.

(Preliminary step) First, make sure that the "Analysis ToolPak" is installed.
Under "Tools" is the option "Data Analysis" present?
If yes – ToolPak is installed.
If no – select "Add-ins."
Check the boxes entitled "Analysis ToolPak" and "Analysis ToolPak – VBA" and click "OK". This will install the "Data Analysis ToolPak."

(1.) Under "Tools" select "Data Analysis"
In the window that appears select "ANOVA: One factor" and click "OK."
(2.) Using your mouse highlight the cells containing the data.
(3.) Select "Columns" if each treatment is its own column or "Row" if each treatment is its own row.
(4.) Set your level of significance. (The default is 5% or 0.05.)
(5.) Click "OK" and the ANOVA output will appear on a new worksheet.

ANOVA Results from Excel:

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	3	2000	666.6667	972.3333

Column 2	3	1421	473.6667	2417.333
Column 3	3	1342	447.3333	1737.333

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	86049.55556	2	43024.78	25.17541	0.001207	5.143249
Within Groups	10254	6	1709			
Total	96303.55556	8				

The results under the heading “SUMMARY” simply provides you with summary statistics for each of your samples. The results of the ANOVA test are provided under the heading “ANOVA.” Comparing these figures with the example above, it should be simple to determine the meaning of the Excel output.

Two-Way Analysis of Variance

Note: Much of the math here is tedious but straightforward. We'll skim over it in class but you should be sure to ask questions if you don't understand it.

I. OVERVIEW.

A. Sometimes a researcher might want to simultaneously examine the effects of two treatments (where both treatments have nominal-level measurement).

EXAMPLES:

✓ the effect of sex and race on wages
✓ the effects of the level of pollution and the level of city services on housing prices
✓ the effects of religion and region on income

To elaborate: with sex and race, we might wonder if
✓ are there differences because of sex alone
✓ are there differences because of race alone
✓ are there differences attributable to particular combinations of sex and race - that is, are there interaction effects? For example, white males, white females, and black males may all have similar wages, but black females could have much lower wages. We'll discuss interaction effects more shortly.

B. Two-Way Anova with a Balanced Design and the Classic Experimental Approach. We can use Analysis of Variance techniques for these and more complicated problems. These techniques can get fairly involved and employ several different options, each of which has various strengths and weaknesses. If this were a psychology class, we might spend a lot more time going over ANOVA, where such techniques are more widely used. But, in Sociology, we are much more likely to use regression and other techniques for our advanced work. Therefore, for our purposes, I will primarily focus on the special case of *balanced designs* (this is also what Hays, Harnett and other texts focus on). In a balanced design, *all cell frequencies are equal*, i.e. the number of observations in each combination of treatments is the same. So, for example, there would be 5 white males, 5 black males, 5 white females, and 5 black females. Balanced designs are unlikely in survey research but they are quite common (and often encouraged) in experimental studies. Equal cell frequencies make it easier to disentangle the effects of the row and column variables (e.g. sex and race) and also minimizes the effect of non-homogenous population variances if they exist.

In addition, I'll note that several programs give you various options for the "Method" to use for Anova. If the design is balanced, I don't think it matters what method you use. But, if you choose what SPSS calls the *Classic Experimental Approach*, many of the formulas that follow will be valid even when the design is not balanced. The *Regression Approach* and the *Hierarchical Approach* are other options (and several other options, with varying names, are also listed in different procedures). The SPSS manual and other sources have more information if you find yourself needing to know about these.

As noted below, these assumptions are not required for everything we will be talking about. These assumptions will affect how computations are done with the raw data but, once that is done, the hypothesis testing procedures will be largely the same. Ergo, the most critical parts of our discussion will apply even when designs are not balanced.

C. The model. When we have 2 treatments, the model can be written as

$$y_{ijk} = \mu + \tau_j + \lambda_k + (\tau\lambda)_{jk} + \varepsilon_{ijk}$$

where μ = the grand mean, τ_j is the treatment effect for the jth category of the row variable, λ_k is the treatment effect for the kth category of the column variable, $(\tau\lambda)_{jk}$ is the interaction effect for the combination of the jth row category and the kth column category.

EXAMPLE: Suppose the overall average income is $20,000, the average black income is $15,000, the average female income is $17,000, and the average black woman's income is $10,000. This means that μ = $20,000, τ_B = -$5,000, λ_W = -$3,000, $(\tau\lambda)_{BW}$ = -$2,000.

D. As before, we want to partition the variance. Note that

$$s_y^2 = \frac{\Sigma\Sigma\Sigma(y_{ijk} - \bar{y})^2}{N - 1} = \frac{Total\ SS}{N - 1} = \frac{TSS}{N - 1} = MS\ Total$$

Further, note that

Component	Description
$(y_{ijk} - \bar{y}) =$	Deviation of the individual score from the overall mean
$(y_{ijk} - \bar{y}_{jk})$	Deviation of the individual score from the group mean, i.e. $\hat{\varepsilon}_{ijk}$
$+(\bar{y}_j - \bar{y})$	Deviation of the jth row's mean from the overall mean, i.e. $\hat{\tau}_j$
$+(\bar{y}_k - \bar{y})$	Deviation of the kth column's mean from the overall mean, i.e. $\hat{\lambda}_k$
$+(\bar{y}_{jk} - \bar{y}_j - \bar{y}_k + \bar{y})$	Deviation of "combination" mean from row and column means; the interaction, i.e. $(\hat{\tau}\hat{\lambda})_{jk}$

Note that we are using the same trick we did before of adding and then subtracting the same terms.

Hence, $\Sigma\Sigma\Sigma(y_{ijk} - \bar{y})^2$ can be broken out as follows (any seemingly omitted terms conveniently work out to be zero):

$$\Sigma\Sigma\Sigma(y_{ijk} - \overline{y}_{jk})^2 = \Sigma\Sigma\Sigma\,\hat{\varepsilon}_{ijk}^2 = SS\ Error,$$

$$d.f. = N - JK$$

This is analogous to SS Within from 1-way ANOVA. This represents the deviation of individuals from the means of others who have the same value on the row and column variables (e.g. are of the same sex and race); that is, this represents the component of the scores that cannot be accounted for by group membership. The d.f. arise from the fact that there are N cases, and J*K means have to be estimated. Also,

$$\Sigma\Sigma\Sigma(\overline{y}_j - \overline{y})^2 = \Sigma\Sigma\Sigma\,\hat{\tau}_j^2 = SS\ Rows,$$

$$d.f. = J - 1$$

$$\Sigma\Sigma\Sigma(\overline{y}_k - \overline{y})^2 = \Sigma\Sigma\Sigma\,\hat{\lambda}_k^2 = SS\ Columns,$$

$$d.f. = K - 1$$

$$\Sigma\Sigma\Sigma(\overline{y}_{jk} - \overline{y}_j - \overline{y}_k + \overline{y})^2 = \Sigma\Sigma\Sigma(\hat{\tau}\hat{\lambda})_{jk}^2 = SS\ Interaction,$$

$$d.f. = (J - 1)(K - 1)$$

Other useful partitionings include

$$SS\ Main = SS\ Total - SS\ Interaction - SS\ \mathrm{Re}sidual$$

$$d.f. = J + K - 2$$

Note also that, *when all cell frequencies are equal*, i.e. the number of observations in each combination of treatments is the same,

SS Main = SS Columns + SS Rows.

This will not necessarily be true otherwise. The fact that it is true in a balanced design is one of its main advantages.

Another useful partitioning is

$$SS\ Cells = SS\ Explained = SS\ Main + SS\ Interaction = SS\ Total - SS\ Error$$

$$d.f. = JK - 1$$

When all cell frequencies are equal,

SS Cells = SS Columns + SS Rows + SS Interaction.

Finally, note that,

$$Total\ SS = SS\ Main + SS\ Interactions + SS\ Error = SS\ Explained + SS\ Error$$

$$d.f. = J - 1 + K - 1 + JK + 1 - J - K + N - JK = N - 1$$

Again, *when all cell frequencies are equal,*

Total SS = SS Columns + SS Rows + SS Interaction + SS Error.

E. *When doing statistical inference, we assume that*

- ✓ for each treatment combination JK, the random error terms ε_{ijk} are ~ $N(0, \sigma^2)$; the variance σ^2 is the same for each treatment combination.
- ✓ the random error terms are independent

II. TESTS OF INTEREST:

A. H_0: $(\tau\lambda)_{jk} = 0$ for all j, k
H_A: $(\tau\lambda)_{jk} <> 0$ for at least 1 j, k

This is a test of whether there are any <u>interaction effects</u>; the appropriate <u>test statistic</u> is

$$F_{(J-1)(K-1), N-JK} = \frac{SS\ Interaction/(J-1)(K-1)}{SS\ Error/(N-JK)} = \frac{MS\ Interaction}{MS\ Error}$$

If the null hypothesis is true, F ~ **F**([J - 1][K - 1], N - JK)

B. H_0: $\tau_1 = \tau_2 = ... = \tau_J = 0$
H_A: At least 1 $\tau_j <> 0$

This tests whether there are any row effects. The appropriate test statistic is

$$F_{J-1,N-JK} = \frac{\textit{SS Rows/(J - 1)}}{\textit{SS Error/(N - JK)}} = \frac{\textit{MS Rows}}{\textit{MS Error}}$$

If the null hypothesis is true, F ~ **F**([J - 1], N - JK)

C. H_0: $\lambda_1 = \lambda_2 = ... = \lambda_K = 0$
H_A: At least 1 λ_k <> 0

This tests whether there are any column effects. The appropriate test statistic is

$$F_{K-1,N-JK} = \frac{\textit{SS Columns/(K - 1)}}{\textit{SS Error/(N - JK)}} = \frac{\textit{MS Columns}}{\textit{MS Error}}$$

If the null hypothesis is true, F ~ **F**([K - 1], N - JK).

NOTE: The last two tests are primarily of interest if you conclude that interaction effects are not significant. If, on the other hand, you conclude that the interaction effects do not equal zero, then you know both treatments (i.e. the row and column effects) are significant.

D. H_0: All τ's and λ's = 0
H_A: At least one τ or λ does not equal 0

This tests whether any of the main effects (i.e. row or column effects; or, non-interaction effects) are nonzero. The appropriate test statistic is

$$F_{J+K-2,N-JK} = \frac{\textit{SS Main/(J + K - 2)}}{\textit{SS Error/(N - JK)}} = \frac{\textit{MS Main}}{\textit{MS Error}}$$

If the null hypothesis is true, F ~ **F**([J + K - 2], N - JK).

E. H_0: All τ's, λ's, and $(\tau\lambda)$'s = 0
H_A: At least one τ, λ, or $(\tau\lambda)$ does not equal 0

This tests whether there are any effects at all. If the null hypothesis is true, then every cell in the table will have the same true mean. The appropriate test statistic is

$$F_{JK-1,N-JK} = \frac{\textit{SS Cells/(JK - 1)}}{\textit{SS Error/(N - JK)}} = \frac{\textit{MS Cells}}{\textit{MS Error}}$$

If the null hypothesis is true, F ~ **F**([JK - 1], N - JK).

III. ROW, COLUMN, AND INTERACTION EFFECTS – EXAMPLES

What are interaction effects? Here are some substantive examples:

✓ Medicines A and B may have no effect when either is taken alone. But, the two together may have an effect. "The whole is different from the sum of the parts."

✓ Another example: we might find that greater income leads to greater fertility for those who want children, and lower fertility for those who do not want children. We say that the effect of income is dependent on desires, or that desires and income interact in determining fertility.

✓ Good teachers and small classrooms might both encourage learning. A good teacher in a small classroom might be especially effective. The whole is greater than the sum of the parts.

Following are hypothetical 2-way ANOVA examples. The dependent variable is income (in thousands of dollars), the row variable is gender (Male or Female), the column variable is type of occupation (A, B, or C). Unless otherwise stated, assume that frequencies are equal for all cells.

1. Row (Gender) effects only.

	Occ A	Occ B	Occ C	
Male	$\mu_{MA} = 18$ $\tau\lambda_{MA} = 0$	$\mu_{MB} = 18$ $\tau\lambda_{MB} = 0$	$\mu_{MC} = 18$ $\tau\lambda_{MC} = 0$	$\mu_M = 18$ $\tau_M = 2$
Female	$\mu_{FA} = 14$ $\tau\lambda_{FA} = 0$	$\mu_{FB} = 14$ $\tau\lambda_{FB} = 0$	$\mu_{FC} = 14$ $\tau\lambda_{FC} = 0$	$\mu_F = 14$ $\tau_F = -2$
	$\mu_A = 16$ $\lambda_A = 0$	$\mu_B = 16$ $\lambda_B = 0$	$\mu_C = 16$ $\lambda_C = 0$	$\mu = 16$

The 2 rows differ, but the three columns are all the same. Within each occupation, men make $4,000 more on average than do women; each of the three occupations pays equally well.

2. Column (Occupation) effects only.

	Occ A	Occ B	Occ C	
Male	$\mu_{MA} = 12$ $\tau\lambda_{MA} = 0$	$\mu_{MB} = 16$ $\tau\lambda_{MB} = 0$	$\mu_{MC} = 20$ $\tau\lambda_{MC} = 0$	$\mu_M = 16$ $\tau_M = 0$
Female	$\mu_{FA} = 12$ $\tau\lambda_{FA} = 0$	$\mu_{FB} = 16$ $\tau\lambda_{FB} = 0$	$\mu_{FC} = 20$ $\tau\lambda_{FC} = 0$	$\mu_F = 16$ $\tau_F = 0$
	$\mu_A = 12$ $\lambda_A = -4$	$\mu_B = 16$ $\lambda_B = 0$	$\mu_C = 20$ $\lambda_C = 4$	$\mu = 16$

The three columns differ, but the two rows are the same. Occupation C pays better than B and B pays better than A. Within each occupation, however, men and women make the same.

3. Row and column effects.

	Occ A	Occ B	Occ C	
Male	$\mu_{MA} = 14$ $\tau\lambda_{MA} = 0$	$\mu_{MB} = 18$ $\tau\lambda_{MB} = 0$	$\mu_{MC} = 22$ $\tau\lambda_{MC} = 0$	$\mu_M = 18$ $\tau_M = 2$
Female	$\mu_{FA} = 10$ $\tau\lambda_{FA} = 0$	$\mu_{FB} = 14$ $\tau\lambda_{FB} = 0$	$\mu_{FC} = 18$ $\tau\lambda_{FC} = 0$	$\mu_F = 14$ $\tau_F = -2$
	$\mu_A = 12$ $\lambda_A = -4$	$\mu_B = 16$ $\lambda_B = 0$	$\mu_C = 20$ $\lambda_C = 4$	$\mu = 16$

Both the rows and columns differ. Within each occupation, men make $4,000 more on average than women do. Within each gender, those in occupation C average $4,000 more than those in B, and those in B average $4,000 more than those in A.

4. Interaction effects I.

	Occ A	Occ B	Occ C	
Male	$\mu_{MA} = 15$ $\tau\lambda_{MA} = -1$	$\mu_{MB} = 15$ $\tau\lambda_{MB} = -1$	$\mu_{MC} = 21$ $\tau\lambda_{MC} = 2$	$\mu_M = 17$ $\tau_M = 1$
Female	$\mu_{FA} = 15$ $\tau\lambda_{FA} = 1$	$\mu_{FB} = 15$ $\tau\lambda_{FB} = 1$	$\mu_{FC} = 15$ $\tau\lambda_{FC} = -2$	$\mu_F = 15$ $\tau_F = -1$
	$\mu_A = 15$ $\tau_A = -1$	$\mu_B = 15$ $\tau_B = -1$	$\mu_C = 18$ $\tau_C = 2$	$\mu = 16$

Five of the six cells have the same mean. However, for some reason, the combination of males and occupation C results in high male earnings.

5. Interaction effects II - differing magnitudes of effects.

	Occ A	Occ B	Occ C	
Male	$\mu_{MA} = 12$ $\tau\lambda_{MA} = -1$	$\mu_{MB} = 16$ $\tau\lambda_{MB} = -1$	$\mu_{MC} = 26$ $\tau\lambda_{MC} = 2$	$\mu_M = 18$ $\tau_M = 2$
Female	$\mu_{FA} = 10$ $\tau\lambda_{FA} = 1$	$\mu_{FB} = 14$ $\tau\lambda_{FB} = 1$	$\mu_{FC} = 18$ $\tau\lambda_{FC} = -2$	$\mu_F = 14$ $\tau_F = -2$
	$\mu_A = 11$ $\lambda_A = -5$	$\mu_B = 15$ $\lambda_B = -1$	$\mu_C = 22$ $\lambda_C = 6$	$\mu = 16$

Men make more than women, and the advantage is especially great in occupation C. Or, those in occupation C make more than those in other occupations, and the advantage is especially great for men.

6. Interaction effects III - differing directions of effects.

	Occ A	Occ B	Occ C	
Male	μ_{MA} = 18 $\tau\lambda_{MA}$ = +2	μ_{MB} = 16 $\tau\lambda_{MB}$ = 0	μ_{MC} = 14 $\tau\lambda_{MC}$ = -2	μ_M = 16 τ_M = 0
Female	μ_{FA} = 14 $\tau\lambda_{FA}$ = -2	μ_{FB} = 16 $\tau\lambda_{FB}$ = 0	μ_{FC} = 18 $\tau\lambda_{FC}$ = 2	μ_F = 16 τ_F = 0
	μ_A = 16 λ_A = 0	μ_B = 16 λ_B = 0	μ_C = 16 λ_C = 0	μ = 16

In this example, the effect of gender depends on occupation. Males do better than women in Occupation A but worse in occupation C; in Occupation B there is no difference. Or, occupation C is better paying for women but not for men, whereas for occupation A the opposite is true. Note that, if you only looked at the main effects, you would erroneously conclude that gender and occupation have no effects on income, when in reality they do have effects but the effects work in opposing directions.

IV. Computational Procedures - Two-Way Anova – Balanced Designs

Let A = row variable, B = column variable, J = number of categories for A, K = number of categories for B, T_{Aj} = the sum of the scores in group A_j, T_{Bk} = the sum of the scores in group B_k, T_{AjBk} is the sum of the scores for the observations which fall in both groups A_j and B_k (there are J*K of these totals), n_{Aj} = number of observations in group A_j, n_{Bk} = number of observations in group B_k, n_{AjBk} is the number of observations which fall in both groups A_j and B_k. [NOTE: While I will show you how to do the raw data calculations, in practice they are tedious enough that I generally would not expect you to do them by hand, at least on an exam. You should know how to do the other formulas, however, as they show how the different parts of the ANOVA table are related to each other.]

Note that many (albeit not all) of the formulas for raw data calculations and Sums of Squares assume a *balanced design*, i.e. all cell frequencies are equal for each possible combination of values for the row and column variables. Computations are somewhat more complicated when designs are not balanced. *The Mean Square formulas and the F tests are accurate regardless of whether the design is balanced or not.*

Formula	**Explanation**
Raw Data Calculations (Balanced Design)	
(1) = $(\Sigma\Sigma\Sigma y_{ijk})^2/n = N\hat{\mu}^2$	Sum all the observations. Square the result. Divide by the total number of observations.
(2) = $\Sigma\Sigma\Sigma y_{ijk}^2$	Square each observation. Sum the squared observations.
(3) = $\Sigma\ T_{Aj}^2/n_{Aj}$	Add up the values for the observations for group A_1. Square the result. Divide by the number of observations in group A_1. Repeat for each category of A. Add the results for each of the J groups together.
(4) = $\Sigma\ T_{Bk}^2/n_{Bk}$	Add up the values for the observations for group B_1. Square the result. Divide by the number of observations in group B_1. Repeat for each category of B. Add the results for each of the K groups together.
(5) = $\Sigma\Sigma\ T_{AjBk}^2/n_{AjBk}$	Add up the values for the observations which fall in both group A_1 and B_1. Square this value, and divide by n_{A1B1}. Repeat for each of the J*K combinations, and sum the results.

Sums of Squares Calculations (Balanced Design)	
SS Total = (2) - (1)	Total sum of squares
SS Rows = (3) - (1)	Row sum of squares. This is also sometimes called SS_A.
SS Columns = (4) - (1)	Column sum of squares. Also called SS_B.
SS Interaction = (5) + (1) - (3) - (4) = SS Total - SS Rows - SS Columns - SS Error = SS Total – SS Main – SS Error	Interaction sum of squares. Also called SS_{AB}. It may be easier to use the second formula.
SS Error = (2) - (5) = SS Total - SS Cells	Error sum of squares. It is analogous to SS Within in one-way ANOVA. Also called SS Residual.
SS Main = (3) + (4) – [2 * (1)] = SS Columns + SS Rows = SS Total – SS Error – SS Interaction	Main effects Sum of Squares. Also called SS_{A+B}
SS Cells = (5) - (1) = SS Main + SS interaction = SS Total - SS Error.	This is analogous to SS Between in one-way ANOVA. Also called SS Explained.
Mean Square Calculations (Balanced or unbalanced)	
MS Total = s^2 = SS Total/(n-1)	Remember that MS Total = s^2
MS Rows = SS Rows/(J-1)	Also called MS_A.
MS Columns = SS Columns/(K-1)	Also called MS_B.
MS Interaction = SS Interaction/((J-1)(K-1))	Also called MS_{AB}
MS Main = SS Main/(J + K - 2)	Also called MS_{A+B}
MS Cells = SS Cells/((J*K)-1)	Also called MS Explained.
MS Error = SS Error/ (n - J*K)	Also called MS Residual.

Possible F Tests (Balanced or unbalanced):	
MS Rows/MS Error	Do means differ across categories of the row variable, i.e. do tau’s differ? d.f. = J-1, n-J*K
MS Columns/MS Error	Do means differ across categories of the column variable, i.e. do lambdas differ? d.f. = K-1, n-J*K
MS Interaction/MS Error	Do any of the interaction effects differ from zero? d.f. = (J-1)(K-1), n-J*K
MS Main/MS Error	Are any of the row or column effects nonzero? d.f. = J + K - 2, n-J*K
MS Cells/MS Error	Are there any differences anywhere across groups? d.f. = (JK-1), N-JK.

An ANOVA table often looks something like this (with the computed values substituted).

Source	SS	D.F.	Mean Square	F
A + B (or Main Effects)	SS Main	J + K - 2	SS Main (J + K - 2)	MS Main MS Error
A (or main effect of A)	SS Rows	J - 1	SS Rows (J - 1)	MS Rows MS Error
B (or main effect of B)	SS Columns	K - 1	SS Columns (K - 1)	MS Columns MS Error
AB (or 2-way interaction)	SS Interaction	(J - 1) * (K - 1)	SS Interaction (J -1) (K - 1)	MS Interaction MS Error
A + B + AB (or explained)	SS Cells	(J * K) - 1	SS Cells (J * K) - 1	MS Cells MS Error
Error (or residual)	SS Error	N - (J * K)	SS Error (N - J * K)	
Total	SS Total	N - 1	SS Total (N - 1)	

V. EXAMPLES.

1. A researcher is interested in differences in income by Region (North, South, East, and West) and Religion (Catholic, Protestant, Other). She draws a sample of ten people for each combination of region and religion. She finds that SS Rows = 200, SS Columns = 170, SS Interaction = 100, and s^2 = 16.81. Construct the Anova Table, and indicate which effects are significant at the .05 level. (NOTE: Region is the row variable.)

Solution. Again the design is balanced. You don't have to do any work with the raw data here; instead, you have to understand how the different parts of the ANOVA table are related to each other. Let us begin with what we are told:

Source	SS	D.F.	Mean Square	F
A + B (or Main Effects)	SS Main =	J + K - 2 =	SS Main = (J + K - 2)	MS Main = MS Error
A (or main effect of A)	SS Rows = **200**	J - 1 =	SS Rows = (J - 1)	MS Rows = MS Error
B (or main effect of B)	SS Columns = **170**	K - 1 =	SS Columns = (K - 1)	MS Columns = MS Error
AB (or 2-way interaction)	SS Intraction = **100**	(J - 1) * (K - 1) =	SS Intrction = (J -1)(K - 1)	MS Intrction = MS Error
A + B + AB (or explained)	SS Cells =	(J * K) - 1 =	SS Cells = (J * K) - 1	MS Cells = MS Error
Error (or residual)	SS Error =	N - (J * K) =	SS Error = (N - J * K)	
Total	SS Total =	N - 1 =	SS Total = **16.81** (N - 1)	

We are also told J = 4 (there are 4 regions), K = 3 (3 religions).

✓ We can deduce that N = J*K*10 = 120.

✓ Recall that s^2 = MS Total, and that MS Total = SS Total/(n-1)
==> SS Total = s^2 * (N - 1) = 16.81 * 119 = 2000.

✓ SS Main is obtained by adding SS Rows + SS Columns = 200 + 170 = 370.

✓ SS Cells is obtained by adding up SS Columns + SS Rows + SS Interactions
= 200 + 170 + 100 = 470.

✓ SS Error is obtained by computing SS Total - SS Cells = 2000 - 470 = 1530.

✓ The remaining quantities in the table are obtained by filling in the appropriate values for the formulas. Hence, we get (* = significant at the .05 level):

Source	SS	D.F.	Mean Square	F
A + B (or Main Effects)	SS Main = 370	J + K - 2 = 5	SS Main = 74.00 (J + K - 2)	MS Main = 5.22* MS Error
A (or main effect of A)	SS Rows = 200	J - 1 = 3	SS Rows = 66.67 (J - 1)	MS Rows = 4.71* MS Error
B (or main effect of B)	SS Columns = 170	K - 1 = 2	SS Columns = 85.00 (K - 1)	MS Columns = 6.0* MS Error
AB (or 2-way interaction)	SS Intraction = 100	(J - 1) * (K - 1) = 6	SS Intrction = 16.67 (J -1)(K - 1)	MS Intrction = 1.18 MS Error
A + B + AB (or explained)	SS Cells = 470	(J * K) - 1 = 11	SS Cells = 42.73 (J * K) - 1	MS Cells = 3.02* MS Error
Error (or residual)	SS Error = 1530	N - (J * K) = 108	SS Error = 14.17 (N - J * K)	
Total	SS Total = 2000	N - 1 = 119	SS Total = 16.81 (N - 1)	

Conclusion. Interaction effects are not significant, other effects are.

2. A consumer research firm wants to compare three brands of radial tires (X, Y, and Z) in terms of tread life over different road surfaces. Random samples of four tires of each brand are selected for each of three surfaces (asphalt, concrete, gravel). A machine that can simulate road conditions for each of the road surfaces is used to find the tread life (in thousands of miles) of each tire. Construct an ANOVA table and conduct F-tests for the presence of nonzero brand effects, road surface effects, and interaction effects.

Surface/ Brand	X	Y	Z
Asphalt	36, 39, 39, 38	42, 40, 39, 42	32, 36, 35, 34
Concrete	38, 40, 41, 40	42, 45, 48, 47	37, 33, 33, 34
Gravel	34, 32, 34, 35	34, 34, 30, 31	36, 35, 35, 33

Solution. I’ll show you how to work this by hand (just in case your life ever depends on it) although on an exam I’d be more likely to give you something like problem 1 and/or give you finished results and ask you to interpret them. More critically, I’ll show you how to do this in SPSS.

Note that the design is balanced. Let A = Road surface, B = Brand. HINT: It is legitimate to subtract a constant from EVERY observation. This will not affect any of the values in the ANOVA table, and it often makes the calculations simpler. I will subtract 30 from each observation, yielding the following table:

Surface/ Brand	X	T_{AjBk}	Y	T_{AjBk}	Z	T_{AjBk}	T_{Aj}
Asphalt	6 9 9 8	32	12 10 9 12	43	2 6 5 4	17	92
Concrete	8 10 11 10	39	12 15 18 17	62	7 3 3 4	17	118
Gravel	4 2 4 5	15	4 4 0 1	9	6 5 5 3	19	43
T_{Bk}		86		114		53	253

(1) = $(\Sigma\Sigma\Sigma y_{ijk})^{2}/n = 253^2/36 = 1778.03$
(2) = $\Sigma\Sigma\Sigma y_{ijk}^{2} = 6^2 + 9^2 + 12^2 + ... + 3^2 = 2451$
(3) = $\Sigma\ T_{Aj}^{2}/n_{Aj} = 92^2/12 + 118^2/12 + 43^2/12 = 2019.75$
(4) = $\Sigma\ T_{Bk}^{2}/n_{Bk} = 86^2/12 + 114^2/12 + 53^2/12 = 1933.42$
(5) = $\Sigma\Sigma\ T_{AjBk}^{2}/n_{AjBk} = 32^2/4 + 39^2/4 + ... + 19^2/4 = 2370.75$

SS Total = (2) - (1) = 2451 - 1778.03 = 672.97
SS Rows = (3) - (1) = 2019.75 - 1778.03 = 241.72
SS Columns = (4) - (1) = 1933.42 - 1778.03 = 155.39
SS Interaction = (5) + (1) - (3) - (4) =
2370.75 + 1778.03 - 2019.75 - 1933.42 = 195.61
SS Main = SS Rows + SS Columns = 397.11
SS Cells = (5) - (1) = 592.72
SS Error = (2) - (5) = 80.25

ANOVA TABLE:

SOURCE	**SS**	**D.F.**	**MEAN SQUARE**	**F**
A + B	397.11	4	99.28	33.43*
A	241.72	2	120.86	40.69*
B	155.39	2	77.70	26.16*
AB	195.61	4	48.90	16.46*
A+B+AB	592.72	8	74.09	24.95*
Error	80.25	27	2.97	
Total	672.97	35	19.23	

* = significant at the .05 level.

NOTE:

- To test for the presence of nonzero road effects, the degrees of freedom = 2,27 and we accept H_0 if F $\leq$ 3.34.
- To test for the presence of nonzero brand effects, d.f. = 2,27 and we accept H_0 if F $\leq$ 3.34.
- To test for the presence of nonzero interaction effects, d.f. = 4,27 and we accept H_0 if F $\leq$ 2.72.
- To test for the presence of any nonzero effects, d.f. = 8, 27 and we accept H_0 if F $\leq$ 2.21.

SPSS Solution. In SPSS, the ANOVA command can be used for 2-way ANOVA problems. Alas, you have to enter the syntax directly – you can't do it with the pull-down menus – but it isn't too hard. If you are bound and determined to use the pull-down menus, you can use the UNIANOVA routine – which I personally find a little confusing but I haven't used it very much. To use UNIANOVA, select ANALYZE/ GENERAL LINEAR MODEL/ UNIVARIATE. Here is how you can work the above problem using the ANOVA routine.

```
DATA LIST FREE / Surface Brand Treadlif.
BEGIN DATA.
1 1 36
1 1 39
1 1 39
1 1 38
1 2 42
1 2 40
1 2 39
1 2 42
1 3 32
1 3 36
1 3 35
1 3 34
2 1 38
2 1 40
2 1 41
2 1 40
2 2 42
2 2 45
2 2 48
2 2 47
2 3 37
2 3 33
2 3 33
2 3 34
3 1 34
3 1 32
3 1 34
3 1 35
3 2 34
3 2 34
3 2 30
3 2 31
3 3 36
3 3 35
3 3 35
3 3 33
END DATA.

VARIABLE LABELS SURFACE 'Type of Surface' BRAND 'Brand of tire'
                TREADLIF 'Tread life (1000s of miles)'.
VALUE LABELS SURFACE 1 'Asphalt' 2 'Concrete' 3 'Gravel'/
```

```
             BRAND 1 'X' 2 'Y' 3 'Z'.
ANOVA /VARIABLES TREADLIF BY SURFACE (1,3) BRAND (1,3)/ Method = Experimental.
```

ANOVA

ANOVA[a]

			Experimental Method				
			Sum of Squares	df	Mean Square	F	Sig.
TREADLIF Tread life (1000s of miles)	Main Effects	(Combined)	397.111	4	99.278	33.402	.000
		SURFACE Type of Surface	241.722	2	120.861	40.664	.000
		BRAND Brand of tire	155.389	2	77.694	26.140	.000
	2-Way Interactions	SURFACE Type of Surface * BRAND Brand of tire	195.611	4	48.903	16.453	.000
	Model		592.722	8	74.090	24.928	.000
	Residual		80.250	27	2.972		
	Total		672.972	35	19.228		

a. TREADLIF Tread life (1000s of miles) by SURFACE Type of Surface, BRAND Brand of tire

VI. N-WAY ANOVA.

It is also possible to address problems where there are more than 2 treatments, e.g. look at the effect of race, sex and religion on income. Things start to get more complicated, of course, but it can be done. Particularly confusing is the fact that you can have 3-way and higher interactions, and it can be difficult to interpret what these mean.

VII. ANALYSIS OF COVARIANCE.

Finally, I'll just briefly note that sometimes problems involve "treatments" (or independent variables) that have both nominal and interval-level measurement. For example, we might be interested in the effects of sex, race, and years of education on income. One way to do this is through Analysis of Covariance. In ANCOVA, continuous variables (in this case education) are referred to as covariates. However, such problems can also be addressed via regression techniques, and since that is the more common strategy in Sociology that is where we will focus our attention. But, if you ever find yourself reading a lot of work in psychology or education or related fields, you may come across references to ANCOVA.